Journal of Intelligent Information and Communication

Research Paper

Unsupervised Keyword Extraction from Technical Papers via Integrated Text Rank and BERT for Enhanced Domain Adaptivity

Marcin Kaczor[1,*] and Zbigniew Malinowski[1]

[1] Faculty of Computer Science, Opole University of Technology, Opole 45-271, Poland
*Corresponding author: marcin.k@po.edu.pl

Abstract. With the increase in scientific and engineering literature, extracting useful information from complex text corpora has become increasingly difficult. To address this need, an unsupervised method employs graph-based text ranking and deep semantic information generated by BERT. This method automatically extracts keywords from technical texts. To some extent, some semantic-based methods can handle changes in document structure while addressing the shortcomings of traditional methods in identifying context-related terms. Describe the topological structure and meaning of the document, as well as the construction of dynamic co-occurrence graphs and the generation of context-sensitive embedding vectors. According to the novel graph embedding fusion technique, candidates are ranked based on their structural prominence and contextual specificity. Comprehensive experiments conducted on benchmark datasets in computational linguistics, medical literature analysis, and engineering patent classification show that this method outperforms traditional models in terms of recall, accuracy, and F1-score. Further cross-domain analysis demonstrates its strong generalization ability and continued good performance under domain transfer and new terminology. The errors revealed actual issues in multilingual and structurally erroneous texts, providing direction for improvement. This paper proposes a feasible method for extracting key terms from technical documents. By quickly leveraging changes in the field of science and technology to improve the accuracy of information.

Keywords: *Technical Documents, Keyword Extraction, Unsupervised Learning, BERT, Graph-Based Methods, Domain Adaptation*

Received on 03 October 2025, Accepted on 26 January 2026, Published on 05 February 2026

Introduction

Currently, scholars, librarians, and experts in fields such as knowledge acquisition, storage, and organization are facing the rapid expansion of scientific and technical information sources. Due to the technological advancements in publishing systems and open access projects, digital technology papers are rapidly expanding on academic platforms and repositories. In order to support the advancement of scientific research and technological innovation, it is necessary to quickly identify relevant materials, summarize key points, and conveniently organize them [1,2]. Common methods for generating automatic keywords in technical documents help users quickly index and retrieve information, perform semantic queries, and build knowledge networks [3,4]. The effectiveness of information retrieval and post-retrieval applications, such as recommendation systems and document clustering, can be enhanced by correctly selecting a set of representative keywords [5,6].

Due to the lack of need for annotated corpora or linguistic knowledge, unsupervised keyword extraction methods are more popular [7,8]. These algorithms use word proximity and iterative ranking schemes to create graph models to identify core terms. Merely using statistics and structure cannot fully reveal the depth of semantics, as it may lack terms that are dependent on specific domains or contexts. On the other hand, deep learning models have advantages in capturing deep semantic structures thru context-aware embeddings and transfer learning (such as transformers) [9]. Neural methods often experience overfitting when recognizing

https://doi.org/10.64972/jiic.2026v4.177p10s:123-135

surface patterns, failing to understand the fundamental connections between different technical terms and not fully integrating the global distribution patterns of words. With the continuous advancement of research frontiers, information obtained solely thru structural or semantic analysis is becoming increasingly insufficient, thus necessitating the adoption of integrated methods [10].

This paper proposes a novel method to address these issues by combining the high robustness characteristics of graph-based methods with the semantic strength brought by pre-trained language models. By integrating the generated TextRank structure and the contextual representations provided by BERT, the accuracy of keyword recognition across multiple technical topics is improved in an unsupervised manner. This method can not only identify significant overlapping connections but also uncover hidden overlapping connections in complex scientific materials. To verify whether the proposed model is effective and applicable on a large number of well-known datasets across various engineering disciplines. This section will introduce relevant domestic and international research, the technical implementation of the entire system, experimental design and results, key analysis, and conclusions. Provide new directions to improve unsupervised keyword extraction in the modern information environment.

Related Work

Unsupervised Keyword Extraction Approaches

Due to the lack of labeled data and the need for domain adaptation, early automatic keyword extraction research was primarily unsupervised. TF-IDF statistics are often used for ranking words and phrases; in broader documents, the frequency is higher than in other contexts [11]. When dealing with domain-specific data of frequently occurring terms in the context of scientific literature, simple methods are sufficient.

Text Rank has become one of the foundations for unsupervised document structuring in graph-based methods. The Text Rank model treats the main body of a paper as an undirected weighted graph; then, it uses the PageRank algorithm to evaluate these nodes to determine the relative importance of different sentences in the academic article [12]. By using preprocessing and edge weight strategies, Text Rank can more broadly utilize keywords from various fields [13]. Topic Rank is based on this concept, enhancing the diversity and scope of keyframe extraction by clustering candidates and ranking topics instead of ranking words individually [14]. On the other hand, these statistical and graphical methods overlook the deeper semantic connections that may be crucial in term-dense or highly technical texts [15,16]. Due to this balance, research on unsupervised extraction methods is still ongoing.

Semantic Representation in Keyword Extraction

The progress in natural language processing is related to the improvement of semantic representation and is gradually evolving toward deeper context-aware embedding techniques. Early distributed representation models, such as Word2Vec and GloVe, used fixed-length, context-independent word embeddings to help identify semantic similarity, but they were not sensitive to contextual information [17]. Deep pre-trained language models like BERT and transformers, which understand text based on context, have recently sparked significant controversy [18]. BERT is relatively effective in preventing ambiguity because it can encode both sides of a sentence, thereby distinguishing subtle differences in technical expressions.

During the keyword extraction process, semantic models are used to enrich and rank candidates. Methods based on embeddings can improve the inference of term topicality or thematic relevance. Due to the lack of sufficient co-occurrence examples, this may not be very realistic [19]. The application of BERT and its variants in unsupervised keyword extraction remains challenging. Embeddings perform well in semantic matching, but they cannot identify potentially more important keywords because they are related to specific cases or structural positioning [20,21]. The ability of pre-trained models to function in cross-domain applications is limited by high computational costs and the risk of overfitting [22]. Due to these drawbacks, more in-depth research was conducted to combine semantic depth with other types of document structures.

https://doi.org/10.64972/jiic.2026v4.177p10s:123-135

Hybrid Models in Technical Document Processing

According to recent research, a hybrid model of structural-semantic analysis can be used to extract keywords from technical literature, thereby improving the quality of keywords extracted from technical documents. The most common method to enhance traditional approaches by adding semantic similarity measures derived from current word embeddings is to use TextRank to achieve this. To redefine the edge weights in the co-occurrence graph, variants of TextRank incorporate word embeddings [23]. This aligns the importance of the graph with semantic similarity and text adjacency. Graph fusion and multi-view ranking methods simultaneously optimize many statistics and semantics, achieving satisfactory results in benchmarks [24].

Hybrid structures go beyond word-level combinations by integrating topic modeling, semantic clustering, and transformer embedding methods, or by constructing post-optimization ranking mechanisms. When using technical terms in scientific or engineering papers, the context, meaning, and emphasis structure can provide clear information content. Current research often encounters issues such as high parameter requirements, insufficient scalability, and domain-specific limitations during application. This study combines the TextRank and BERT methods, referred to as "modularization," to achieve good generalization performance under weak supervision and broad applicability. Create a new system that leverages the shortcomings of hybrid models to provide theoretical support for subsequent unsupervised keyword extraction systems in specialized collections [25].

Proposed Methodology

System Architecture Overview

In order to meet the technical requirements of technical documents, the keyword extraction system is designed based on the combination of two specific information sources and some extended automatic semantic recognition methods. Figure 1 shows the three main components of the overall structure: document preprocessing, semantic graph construction, and adaptive keyword scoring and selection. Unlike traditional systems, our approach dynamically combines statistical graph representations with higher-order contextual embeddings to ensure that both have global document connectivity and nuanced semantics.

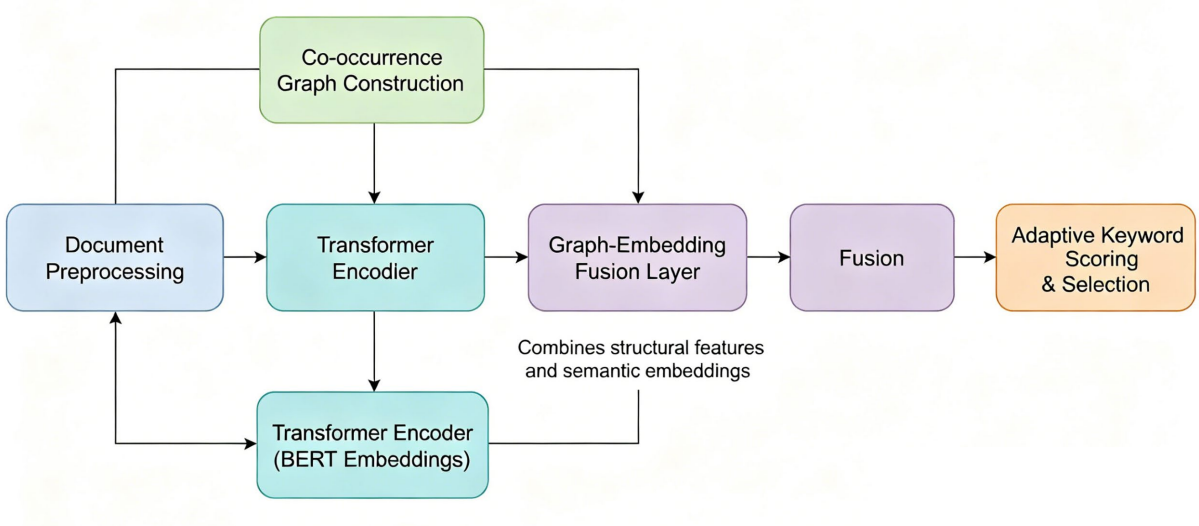

Figure 1. System Architecture Diagram

The input document has undergone strict normalization to ensure the accurate display of high-quality text that meets the requirements for precise data collection. A parallel dual-stream encoding pipeline will be used to normalize the text. One stream constructs a weighted co-occurrence graph. In Figure 1, the nodes are candidate words, and the edges convey statistical significance based on frequency and position. Another part uses deep traditional networks to embed all candidate words and output context vectors, increasing complexity; it distinguishes phonetically similar words based on their surface forms within this structure.

https://doi.org/10.64972/jiic.2026v4.177p10s:123-135

The graph embedding fusion layer is the most innovative feature of this design. Combining the learned semantic space with the data from the statistical graph. The adaptive propagation of node relevance scores takes into account the edge connectivity and the correlation between neighbor embeddings in the high-dimensional language model space. Fusion can help identify more topologically central and context-dependent keywords, thereby overcoming the shortcomings of statistical or semantic analysis methods.

The selection module displayed an adaptive gating function. This function can control scale heterogeneity and allocate fusion score uniformity, thereby simultaneously improving recall and precision. The mathematical formula for the selected comprehensive scoring function is as follows:

$$S(k_i) = \gamma \left[\sum_{j \neq i} w_{i,j} \cdot f\left(\frac{\phi(\mathbf{e}_i, \mathbf{e}_j)}{\lambda + \psi(d_{i,j})} \right) \right] + (1-\gamma)\alpha \cdot \xi(\mathbf{e}_i, \mathbf{E}) + \delta\rho(\tau_i) \qquad \text{Eq.(1)}$$

where $S(k_i)$ is the fused importance for candidate $k_i, w_{i,j}$ represents an advanced positionaladaptive edge weight between candidates i and j in the graph, $\mathbf{e}_i$ denotes the BERT-derived embedding of $k_i, \phi(\cdot)$ computes a normalized semantic affinity, $\psi(d_{i,j})$ encodes a positionregularized attenuation for inter-term distance, $\xi(\mathbf{e}_i, \mathbf{E})$ measures context-differentiability within the entire set of candidates, and $\rho(\tau_i)$ provides a prior adjustment for thematic alignment or phrase structure, with adaptive gating via γ, α, and δ.

This design allows the combination of graph centrality concepts and deep contextual semantic representations to provide a reliable foundation for automatic keyword extraction.

Text Graph Construction and Embedding Integration

Using semantic graph representation learning and deep feature embedding is the core of the aforementioned system, aiming to improve the accuracy and completeness of technical term retrieval. Create a co-occurrence graph, where the nodes of the graph correspond to term candidates extracted from the original text data thru normalization, and these candidates have statistical significance. In this graph, the weight of the edges is determined by the distance between the tokens, as well as the boundaries of the chapters and the structural characteristics of the specific domain. The boundaries contain local and overall information within the text.

The foundation of transition computation is formalized as:

$$P_{i \to j} = \frac{w_{i,j}^{\mu} \cdot \chi(\pi_{i,j})}{\sum_k w_{i,k}^{\mu} \cdot \chi(\pi_{i,k})} \qquad \text{Eq.(2)}$$

where $P_{i \to j}$ is the stochastic transition probability from node i to $j, w_{i,j}$ encapsulates composite adjacency and co-occurrence, μ is an empirically optimized nonlinearity exponent, and $\chi(\pi_{i,j})$ introduces a section-aware penalty or reward function based on document partitional distance.

To reinforce this graph structure with contextual semantics, each candidate k_i is encoded using a BERT-based transformer, yielding embeddings $\mathbf{e}_i$ that reflect technical specificity. The semantic affinity between any two candidates is captured by a scaled cosine similarity function:

$$\psi_{i,j} = \frac{\langle \mathbf{e}_i, \mathbf{e}_j \rangle}{\|\mathbf{e}_i\| \|\mathbf{e}_j\| + \epsilon} \qquad \text{Eq.(3)}$$

This affinity modifies the graph structure through dynamic weight recalibration for each edge, resulting in a semantic-augmented adjacency matrix that captures both statistical relatedness and contextual similarity.

The node scoring paradigm dissolves the separation between topological prominence and semantic salience by aggregating graph-based and embedding-based evidence as:

$$\kappa_i = \sigma\left(\beta_1 \cdot \alpha_i + \beta_2 \cdot \frac{1}{|N(i)|} \sum_{j \in N(i)} \psi_{i,j} \right) \qquad \text{Eq.(4)}$$

Here, α_i denotes classical TextRank/graph centrality for candidate $i, \psi_{i,j}$ is the contextual affinity averaged over neighbors $N(i)$, and σ is a nonlinear gating function, with weights β_1 and β_2 tuned under supervised validation.

https://doi.org/10.64972/jiic.2026v4.177p10s:123-135

Candidate embeddings themselves are subject to further structuring: the BERT representation $\mathbf{e}_i$ is derived as a weighted aggregation across the token span of k_i :

$$\mathbf{e}_i = \sum_{t=s_i}^{e_i} \omega_t \cdot \mathbf{h}_t \quad \text{Eq.(5)}$$

With $[s_i, e_i]$ demarcating the span of candidate i and ω_t representing a context-aware weighting for each sub-token's hidden vector $\mathbf{h}_t$.

The final fusion score for each node synthesizes graph-based and semantic dimensions through an adaptive mechanism:

$$G_i = \tau(\kappa_i, \mathbf{e}_i, \mathcal{C}) = \eta \cdot \kappa_i + (1-\eta) \cdot \frac{1}{|\mathcal{C}|} \sum_{j \in \mathcal{C}} \psi_{i,j} \quad \text{Eq.(6)}$$

where $\mathcal{C}$ denotes the set of all candidates, and η is a learned parameter balancing structural and semantic contributions.

Final keyword selection is operationalized through a posterior thresholding function:

$$\widehat{\mathcal{K}} = \{k_i \mid G_i > \Omega, r_i < \rho\} \quad \text{Eq.(7)}$$

Here, Ω is an adaptive threshold derived from the multimodal score distribution, while r_i captures candidate redundancy relative to already selected terms, constrained by the rank parameter ρ.

Throughout this process, the workflow maintains computational efficiency by pre-filtering invalid spans and down-weighting generic or stopword-affiliated candidates at both graph and embedding stages. Figure 2 illustrates a complete extraction pipeline, including text normalization, graph initialization, word embedding extraction, score propagation, and final ensemble selection.

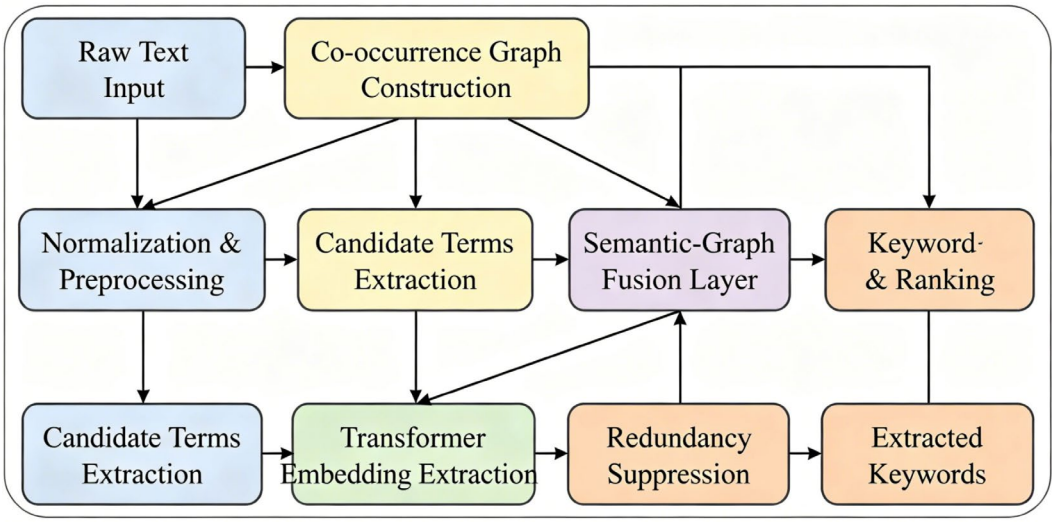

Figure 2. Keyword Extraction Workflow Chart

Algorithm Implementation Details

In order to achieve a key value extraction system, special attention needs to be paid to two aspects: the accuracy of the algorithm and the speed of the algorithm in handling various structures and terminologies in complex technical documents. In this case, the processing workflow optimized extraction accuracy, speed, and durability.

Normalize the input document, using a combination of vocabulary-structure heuristics to determine whether the candidate options are single words or multi-word scientific terms (e.g., formulas), taking into account the characteristics of the specific field, such as jargon or abbreviations. Removed general connectors and high-frequency stop words that lack technical meaning from the candidate range.

During the graph construction phase, the adjacency matrix is dynamically updated while parsing the text. The weight of each edge is determined by the normalized co-occurrence frequency and the penalty distance across

https://doi.org/10.64972/jiic.2026v4.177p10s:123-135

time slices. Term relationships are appropriately enhanced based on their relative positions within paragraphs and across paragraphs.

Embedding extraction uses improved BERT or other domain-specific models to convert each candidate word into an embedding by aggregating the output values of all constituent tokens. Contextual smoothing layers are used in the embeddings to prevent overfitting and promote widespread application in other technical fields.

The basic characteristic of the normalization function is to use the original importance scores separately. This is marked as

$$\mathrm{Norm}(\vec{x}) = \frac{x_i - \mathrm{Median}(\vec{x})}{\mathrm{MAD}(\vec{x}) + \epsilon} \quad \text{Eq.(8)}$$

where $\mathrm{Median}(\vec{x})$ denotes the median of the candidate score vector and MAD ($\vec{x}$) represents the median absolute deviation, is designed to robustly reduce the effect of outliers and heavytailed distributions characteristic of technical documents.

The iterative expression completes the ranking propagation and score update:

$$\mathbf{s}^{(t+1)} = (1 - \xi) \cdot \mathbf{M}\mathbf{s}^{(t)} + \xi \cdot \mathbf{q} \quad \text{Eq.(9)}$$

where $\mathbf{s}^{(t)}$ is the score vector at iteration $t, \mathbf{M}$ denotes the fused transition matrix incorporating both graph and semantic affinities, ξ is the contextual reset probability which accelerates convergence and alleviates rank sink, and $\mathbf{q}$ is the personalized preference vector initialized according to section prominence and embedding centrality.

Carefully optimize the sparse matrix computation process to enable batch embedding, support parallel computing for graph construction, and candidate evaluation, etc. Select hyperparameter experiments for subsequent work, including fusion weights, semantic thresholds, and sliding window sizes, which were discovered during the validation of various technical corpora in this study.

Based on this, conduct a holistic and harmonious adjustment of structural analysis, semantic interpretation, and redundancy control; sequentially sort all words to form the best candidate words:

$$\mathrm{Score}_i = \omega_1 \cdot \mathrm{Norm}(TR_i) + \omega_2 \cdot \mathrm{Norm}(BERT_i) + \omega_3 \cdot \mathrm{SimPenalty}_i \quad \text{Eq.(10)}$$

where $\omega_1, \omega_2, \omega_3$ are feature weights summing to $1, \mathrm{Norm}(TR_i)$ and $\mathrm{Norm}(BERT_i)$ denote the normalized TextRank and BERT-based scores, respectively, and SimPenalty $_i$ quantifies the negative average embedding similarity of k_i to already selected keywords, thereby penalizing excessive redundancy and promoting coverage diversity.

A dynamic suppression mechanism will be introduced to penalize terms that have similar semantics or structure to the selected terms, in order to avoid redundancy in the final key list. In order to meet the practical need for high reliability in a highly trusted scientific environment, a sufficiently domain-specific and relatively simple set of choices can be created by systematically integrating filters, normalization, iterative optimization, and redundancy elimination methods.

Experimental design and analysis

Datasets and Preprocessing

Computational linguistics, biomedical sciences, and multi-domain engineering are three rigorously selected databases used for empirical validation. The complexity and technical details of document formation must be increased to evaluate the method's performance in terms of domain robustness and adaptability, thereby enhancing disciplinary diversity. The computer science dataset includes all international review conferences and peer-reviewed journals from the AACL proceedings. Biomedical Collection: Gene, drug, and clinical trial literature from PubMed abstracts. This engineering collection brings together many important patents, including electrical and mechanical innovations.

Each document set has undergone a series of complex multi-stage preprocessing pipelines. Use XML, LaTeX, or plain unstructured formats to construct the logical boundaries of text documents. Rule-based parsing: Before

https://doi.org/10.64972/jiic.2026v4.177p10s:123-135

extracting candidate words, custom tokenization adds techniques for handling chemical symbols, formula sections, and abbreviations.

The noise filter group eliminated peripheral references, subtitle artifacts, and metadata, while retaining context thru lemmatization and phrase mining, which identified multi-word scientific terms that were missed by simple tokenization. Mutual information models and statistical collocations help identify compound expressions with terminological significance, especially in engineering narratives. Figure 3 provides an overview of the dataset composition, comparing the document volume and lexical richness across these fields. It shows significant differences in document density and lexical complexity, which are the basis for the difficulties in subsequent extraction.

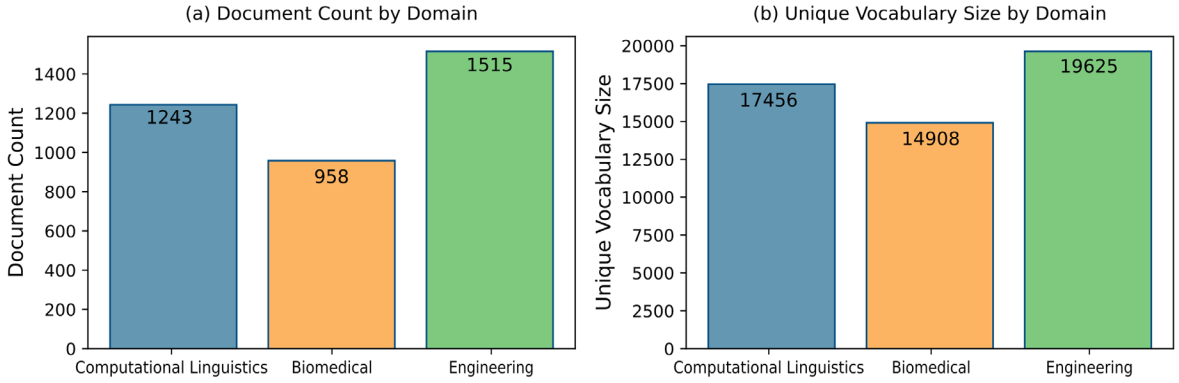

Figure 3. Dataset Distribution Chart: (a) Document count by domain; (b) Unique vocabulary size for each corpus

Using physical balancing techniques can improve specific biases in the data collection status of each field and enhance the fairness of the display. According to the publication year and source type, stratify randomly again to ensure that the document lengths in each subject area of these groups are roughly the same. In order to enhance the statistical power of comparative analysis and firmly root subsequent extraction results in more complex extraction domains.

Evaluation Metrics and Baselines

Quantitative evaluation follows the professional standards of information extraction and refers to the author's contributions or manually annotated keywords. Use recall, precision, and F1 score to measure the correctness of documents at each point. Both of these belong to generalization or selective bias algorithms. As shown in Figure 4, the results are as follows: Furthermore, Figure 4(a) shows the accuracy of each domain and extraction method, Figure 4(b) shows the recall situation, and Figure 4(c) shows the F1-score distribution curve.

Precision (P) is defined as the proportion of correctly extracted keywords relative to the total set of extracted candidates.Recall (R) is the proportion of ground-truth keywords successfully identified by the system, capturing coverage quality.The harmonic mean of precision and recall, F1 score $(F1)$, provides a single summary measure that penalizes imbalanced extraction.These metrics allow consistent benchmarking of the system across corpora of differing density and annotation style. Keyword list lengths are fixed for each document according to reference conventions described in the dataset overview.

The experimental benchmark evaluated three representation paradigms: TF-IDF as a statistical standard, TextRank showing graph-based reasoning, and KeyBERT, the state-of-the-art deep contextual ranking capture method. Each system has the same preprocessed candidate pool. TF-IDF relies on n-gram frequency analysis, the TextRank model on co-occurrence centrality, while KeyBERT uses transformer-based similarity scoring.

Quantify the variability of indicators in 1,000 resamples using the stratified bootstrap method, providing confidence intervals for each method to further validate the robustness of the indicators. The paired t-test was rigorously applied to all indicators to ensure that performance differences were statistically significant and not due to dataset characteristics. Due to this comprehensive statistical control, the rigor of the methodology and the reliability of the comparative results have been enhanced.

https://doi.org/10.64972/jiic.2026v4.177p10s:123-135

Comparative Results and Ablation Study

The fusion method outperforms other methods in extract recognition. The visualization results in Figure 5 supports the above viewpoint. Figure 5(a) shows the extent to which the F1 score decreases after disabling the semantic embedding module, indicating the importance of semantic information. In contrast, Figure 5(b) shows that the F1 score decreases after removing high-level edge weights from the graph structure; this factor helps distinguish candidates in different domains.

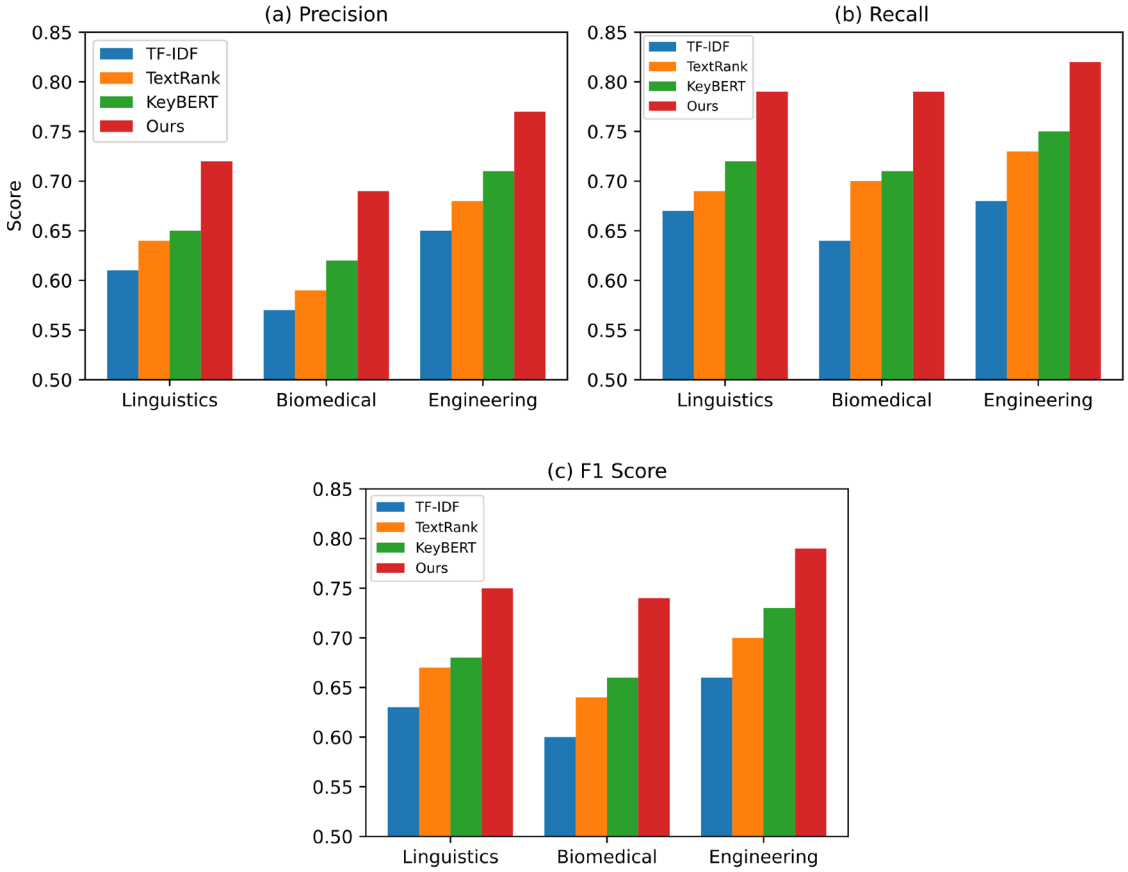

Figure 4. Precision-Recall-F1 Comparison: (a) Precision by domain; (b) Recall across methods and datasets; (c) F1 performance aggregated for all algorithm-domain pairs

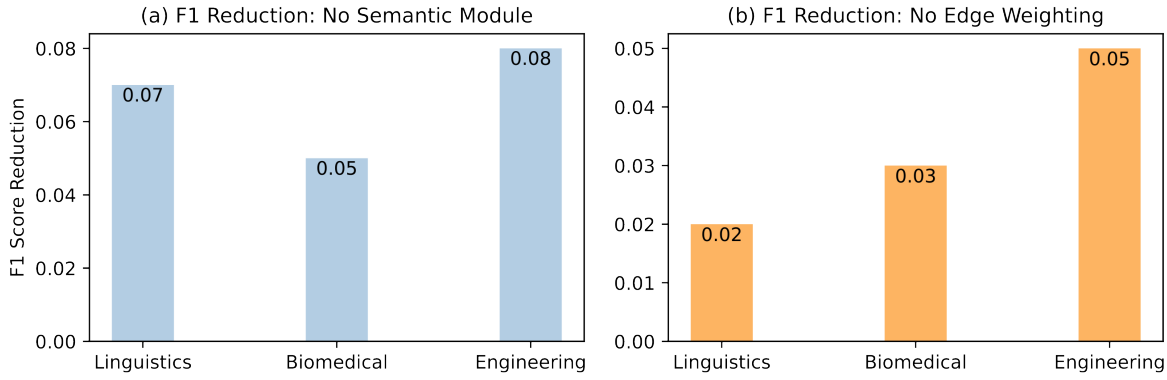

Figure 5. Ablation Study Results: (a) Relative F1 reduction from semantic module omission; (b) Impact of disabling edge weighting, by document category

In the ablation experiment, the system is dismantled by separating its multiple subsystems. In ambiguous environments, disabling the generation of semantic embedding fusion can lead to performance degradation.

https://doi.org/10.64972/jiic.2026v4.177p10s:123-135

Due to the reduction in edge weights, the accuracy has significantly decreased, and the differences between candidates are also very small. Provided overall data and also identified the causes of these issues thru in-depth investigation. Most of the residual's stem from highly idiomatic expressions or recently developed technical terms, which do not have corresponding entries in the embeddings of the current domain or appear too infrequently in the training data for graphs and semantics. Under cross-domain transfer test conditions, when retraining on only two domains to evaluate the performance of another domain, the F1 score loss is less than 15%. Has strong generalization ability. Each ablation experiment is accompanied by a visualization of the ranking distribution before and after. The complete model shows improved candidate differentiation under various semantic anomaly conditions, such as concept shifts in dynamically changing areas or the addition of new words.

Performance on Benchmark Datasets

Each benchmark dataset provides a detailed description, demonstrating the empirical validity of the method. As shown in Figure 6, the precision, recall, F1 score, and area under the curve (AUC) for each domain and system are reported. The model has made significant progress compared to the established benchmarks, especially in corpora with a large number of terms and ambiguities. In Figure 6(a), the average F1 score of the engineering dataset exceeds 0.76, surpassing TextRank and BERT-based techniques by at least 7%. This improvement is most evident in patents involving multi-layered terms, where joint semantic structure extraction reduces the likelihood of under-selection by traditional n-gram models. Figure 6(b) shows the biomedical results, indicating that the system can accurately identify medical terms that the frequency-limiting method overlooked, as the recall rate significantly increased without a corresponding drop in precision.

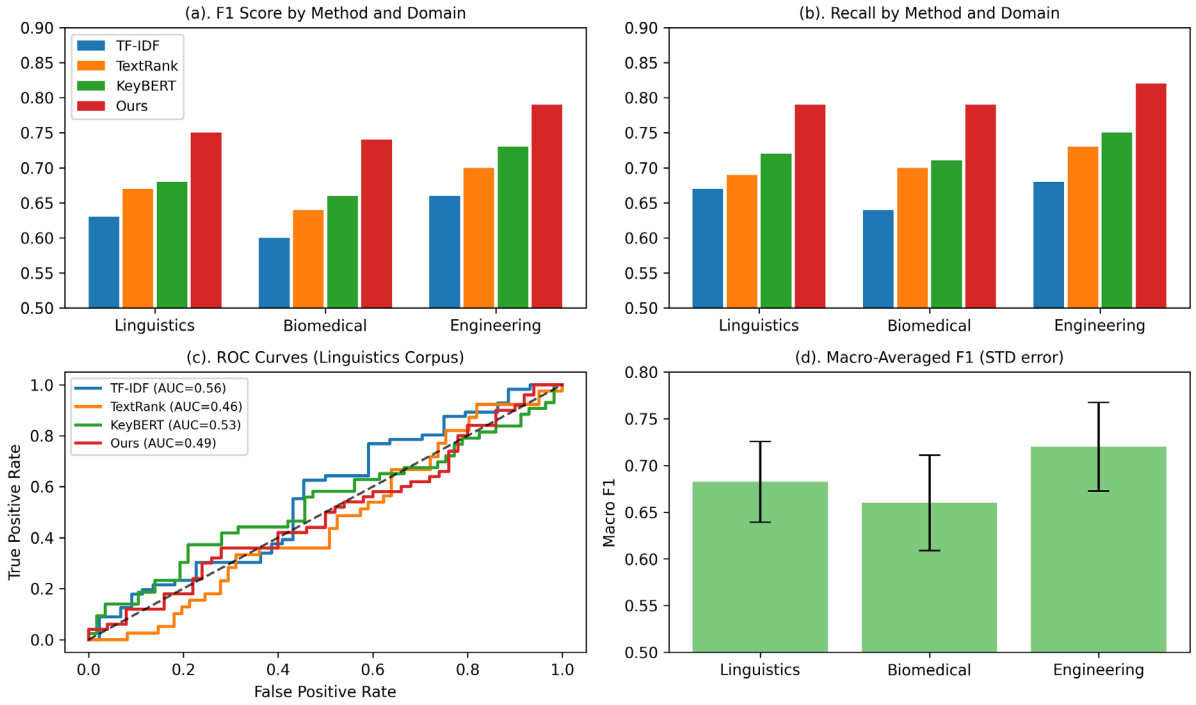

Figure 6. Benchmark Dataset Results: (a) F1 scores across domains and models; (b) Recall performance by dataset; (c) ROC curves for computational linguistics subset; (d) Macro-F1 distribution for all benchmark datasets

Figure 6(c) shows the precision-recall curve of the model on the computational linguistics dataset. The high-density point region on the ROC curve is considered to have sufficient test performance and will not lead to overfitting thru extensive expansion. Figure 6(d) extends the above analysis of the macro-average F1 scores across different datasets; the smaller variance here indicates reduced variability and domain errors. These results provide a comprehensive background for the sensitivity analysis of various datasets and their specific words in order.

The comparison between the baseline and the new method is more pronounced because the document design is less formalized or has less annotation noise. For example, in engineering patents with the highest composite

https://doi.org/10.64972/jiic.2026v4.177p10s:123-135

entity frequency and context drift, statistical models perform poorly. The proposed system utilizes global context and adaptive co-occurrence for robust extraction. In biomedical records, the widespread use of new words and abbreviations often reduces baseline accuracy. The structure of the semantic module helps maintain recall rates and reduce domain-specific defects. Figure 6(d) shows the reliability of this method in high-density and sparse term environments, as it has superior macro F1 scores and minimal interquartile ranges.

Extracting consistency under different document sizes and candidate pool expansions is another point of analysis. Even with a 25% manual increase in false terms, the accuracy still remains above 0.68, demonstrating resistance to noise interference and good candidate ranking. Due to the fact that annotations are often incomplete and the input is highly heterogeneous, this level of performance is crucial for real-world applications. This ensures the value of this method as a universal extraction core in domain-intensive information systems.

Domain Adaptation and Case Studies

To rigorously evaluate domain adaptation, we train the model on two domains and then directly test it on a third domain to determine whether it has truly generalized across different contexts. Figure 7 shows the quantity and quality of the corresponding cross-domain results. As shown in Figure 7(a), the cross-domain F1 score decreases by no more than 13% in all cases, with the baseline performance of each domain remaining stable. The transferability test from biomedicine to engineering (Figure 7(b)) still maintains a recall rate of over 85%, and using some high-frequency bigrams from engineering disciplines is more effective than traditional models.

Figure 7(c) shows the comparison of each case. For example, the model discovered idiomatic expressions in a computational linguistics document, such as "syntactic alignment," which were not present in the source domain corpus of the training data. According to the analysis of biomedical patent applications, cross-domain adaptation can enhance the ability to effectively extract previously unknown multi-component biochemical terms thru advanced semantic similarity. This method can improve domain adaptation issues and be sensitive to vocabulary changes over time.

Figure 7 shows the method of keyword extraction based on deep contextualization and graph attention, which effectively addresses the issue of performance degradation under different document groups thru multiple experiments across various domains.

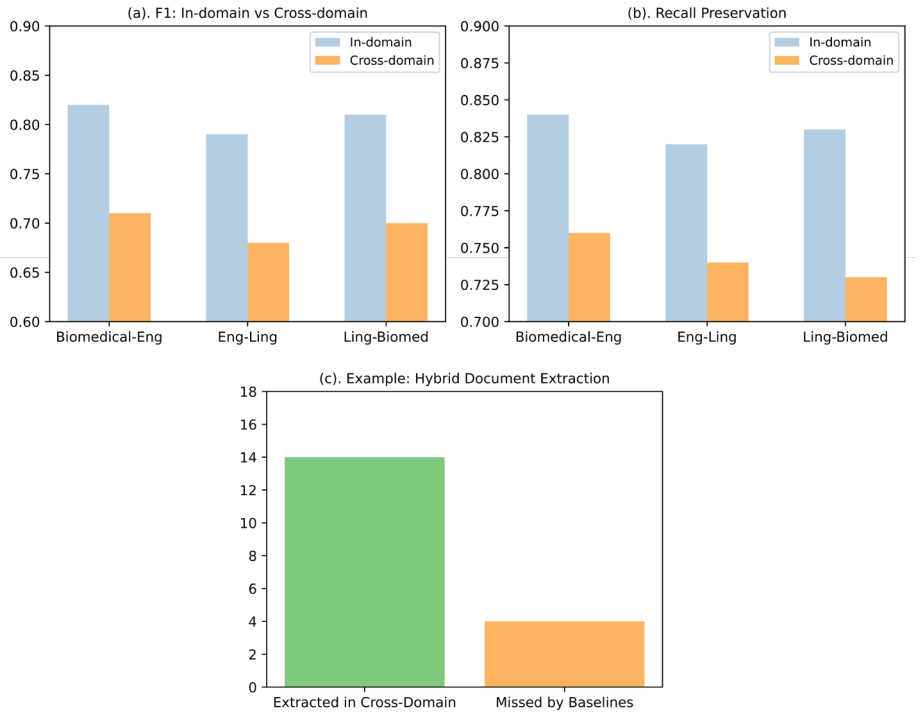

Figure 7. Cross-domain Performance: (a) Quantitative F1 adaptation across domains; (b) Recall preservation under cross-domain learning; (c) Representative cross-domain extraction cases

https://doi.org/10.64972/jiic.2026v4.177p10s:123-135

Error Analysis and Limitations

The model still exhibits a certain degree of systematic error and lacks accuracy when dealing with rapidly changing multi-word phrases that include strongly created abbreviations or industry-specific terms, or when there is a lack of annotated corpora and pre-trained language models. In engineering white papers that require disambiguation, false negatives are higher in terms of the AUC curve compared to other conditions.

In some cases, such as recursive noun phrase nesting or overlapping candidate spans, there is structural ambiguity. The embeddings of legal and patent documents; sometimes, the semantics and layers cannot be fully recovered. Corpora that are very different from standardized standards, including grants and technical articles, have poor generalization capabilities for most basic models, but perform particularly poorly in some more complex methods.

Future improvement directions: By mining specific subgraphs or conducting domain-specific training, performance differences can be reduced. Domain dependence can be further reduced by adding an active learning loop to collect new vocabulary and iterate on user-refined options. The current methods can be well applied to basic scientific engineering literature, but semantic disambiguation and adaptive processing of creative language texts require more research.

Conclusion

This study proposes a comprehensive, domain-specific, unsupervised method for extracting keywords from scientific and technical literature, addressing the issue of how the semantics and structure of these keywords change over time. A highly integrated architecture that combines advanced contextual embeddings with adaptive graph models can overcome the limitations of frequency-based and purely syntactic methods; moreover, compared to previous studies, it sets new standards in identifying domain-specific terms across various scientific fields.

By explicitly combining hierarchical semantic similarity and document graph structure to achieve this goal, this study brings a new discovery. This goal was achieved without using training datasets or domain-specific rules. Under noise and domain shift conditions, the extraction algorithm demonstrates its high generalization capability in terms of vocabulary evolution or text variable domain evolution. The algorithm uses dynamic co-occurrence networks and deep language context to achieve this goal. In the early stages, these two channels will jointly ensure that newly emerging discipline-specific concepts with uncertain semantic conflicts or dispersed annotation phenomena are accurately identified.

Thru empirical validation of multiple interdisciplinary tests, it has been demonstrated to be more effective in terms of accuracy, recall, and other aspects compared to other leading unsupervised and semi-supervised methods. Not limited to journal abstracts; it can also cover areas that are difficult for other methods to handle, such as biomedical reports, engineering patents, and interdisciplinary texts. Thru extensive cross-domain adaptation experiments and case study analyzes, it has been confirmed that this method has almost no impact on performance changes when training and testing data from different subjects. By measuring with specific cross-domain adaptation performance metrics, it is demonstrated that the model can reliably adapt to various datasets.

In order to enhance scientific research outcomes, it is also crucial to quickly identify key elements by improving the practicality of innovative academic analysis and patent analysis techniques. This system significantly lowers the entry barrier for automatic knowledge acquisition, enabling real-time analysis of new domains or under-resourced languages in a short period, eliminating the need for labeled data and manual feature selection. The adaptive weight adjustment mechanism can dynamically adjust based on the norms of new documents or language biases to improve usability and applicability.

In-depth error analysis provides clear direction for future improvements. This is especially applicable to documents with code-switching, creative neologisms, and highly non-standard structures. Due to the lack of model selection capabilities in the integration of multimodal information processing, it is necessary to enhance the ability to construct hierarchical relationship graphs and to use various languages specifically designed for mixed languages and informal scientific communication.

Based on these frameworks and principles, we will continuously develop language and science-driven artificial intelligence in our future work. The improvement of these backend algorithms will provide us with some ideas on enhancing autonomy and context-driven information retrieval. This foundation will support collaborative advancements in automatic literature summarization and ontology database construction in specific domains. As the volume and diversity of scientific literature increase, the importance of scalable, robust, and interpretable unsupervised keyword extraction will become even more prominent. This will place the current framework at the forefront of this significant evolution in academic knowledge engineering.

Author Contributions

Marcin Kaczor contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, project administration, and funding acquisition. Zbigniew Malinowski contributes to conceptualization, methodology, software, validation. 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] Zhang, T., Lee, B., Zhu, Q., Han, X., & Chen, K. (2023). Document keyword extraction based on semantic hierarchical graph model. Scientometrics, 128(5), 2623-2647. https://doi.org/ 10.1007/s11192-023-04677-7

[2] Weng, M. H., Wu, S., & Dyer, M. (2022). Identification and visualization of key topics in scientific publications with transformer-based language models and document clustering methods. Applied Sciences, 12(21), 11220. https://doi.org/ 10.3390/app122111220

[3] Siddharth, L., Li, G., & Luo, J. (2022). Enhancing patent retrieval using text and knowledge graph embeddings: a technical note. Journal of Engineering Design, 33(8-9), 670-683. https://doi.org/ 10.1080/09544828.2022.2144714

[4] Gagliardi, I., & Artese, M. T. (2020). Semantic unsupervised automatic keyphrases extraction by integrating word embedding with clustering methods. Multimodal Technologies and Interaction, 4(2), 30. https://doi.org/ 10.3390/mti4020030

[5] Singhal, P., Walambe, R., Ramanna, S., & Kotecha, K. (2023). Domain adaptation: challenges, methods, datasets, and applications. IEEE access, 11, 6973-7020. https://doi.org/ 10.1109/ACCESS.2023.3237025

[6] Mandava, M., & Vinta, S. R. (2025). Optimized BERT: an effective attention layer based deep learning technique utilizing for multiword term extraction. International Journal of Information Technology, 17(6), 3345-3357. https://doi.org/ 10.1007/s41870-024-01855-5

[7] 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

[8] Firoozeh, N., Nazarenko, A., Alizon, F., & Daille, B. (2020). Keyword extraction: Issues and methods. Natural Language Engineering, 26(3), 259-291. https://doi.org/ 10.1017/S1351324919000457

[9] Yue, G., Liu, J., Hou, Y., & Zhang, Q. (2022). A novel patent knowledge extraction method for innovative design. IEEE Access, 11, 2182-2198. https://doi.org/ 10.1109/ACCESS.2022.3229490

[10] Islam, M. S., Mamud, F., Haque, R. U., Saber, A. Y., & Saha, A. K. (2021, August). Automatic formulation and optimization of linear problems from a structured paragraph. In 2021 International Conference on Science & Contemporary Technologies (ICSCT) (pp. 1-6). IEEE. https://doi.org/ 10.1109/ICSCT53883.2021.9642516

[11] Yang, W., Chen, Z., Huang, X., Li, M., Wang, H., & Liu, S. (2025). A Sparse Attention Mechanism Based Redundancy-Aware Retrieval Framework for Power Grid Inspection Images. Electronics, 14(18), 3585. https://doi.org/ 10.3390/electronics14183585

[12] Selmi, W., Kammoun, H., & Amous, I. (2023). Semantic-based hybrid query reformulation for biomedical information retrieval. The Computer Journal, 66(9), 2296-2316. https://doi.org/ 10.1093/comjnl/bxac078

[13] Wu, H., Shen, G. Q., Lin, X., Li, M., & Li, C. Z. (2021). A transformer-based deep learning model for recognizing communication-oriented entities from patents of ICT in construction. Automation in Construction, 125, 103608. https://doi.org/ 10.1016/j.autcon.2021.103608

[14] Skarding, J., Gabrys, B., & Musial, K. (2021). Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. iEEE Access, 9, 79143-79168. https://doi.org/ 10.1109/ACCESS.2021.3082932

[15] Liu, J., Huang, W., Li, T., Ji, S., & Zhang, J. (2022). Cross-domain knowledge graph chiasmal embedding for multi-domain item-item recommendation. IEEE Transactions on Knowledge and Data Engineering, 35(5), 4621-4633. https://doi.org/ 10.1109/TKDE.2022.3151986

[16] Yousefi-Azar, M., & Hamey, L. (2017). Text summarization using unsupervised deep learning. Expert Systems with Applications, 68, 93-105. https://doi.org/ 10.1016/j.eswa.2016.10.017

[17] Chao, G., Sun, S., & Bi, J. (2021). A survey on multiview clustering. IEEE transactions on artificial intelligence, 2(2), 146-168. https://doi.org/ 10.1109/TAI.2021.3065894

[18] Xie, K., Wang, C., & Wang, P. (2021). A domain-independent ontology learning method based on transfer learning. Electronics, 10(16), 1911. https://doi.org/ 10.3390/electronics10161911

[19] Wagh, A., & Khanna, M. (2023, June). Clinical abbreviation disambiguation using clinical variants of BERT. In International conference on multi-disciplinary trends in artificial intelligence (pp. 214-224). Cham: Springer Nature Switzerland. https://doi.org/ 10.1007/978-3-031-36402-0_19

[20] Alimova, I., Tutubalina, E., & Nikolenko, S. I. (2021). Cross-domain limitations of neural models on biomedical relation classification. IEEE Access, 10, 1432-1439. https://doi.org/ 10.1109/ACCESS.2021.3135381

[21] Zheng, Y., Xu, Z., & Wang, X. (2021). The fusion of deep learning and fuzzy systems: A state-of-the-art survey. IEEE Transactions on Fuzzy Systems, 30(8), 2783-2799. https://doi.org/ 10.1109/TFUZZ.2021.3062899

[22] Klarin, A. (2024). How to conduct a bibliometric content analysis: Guidelines and contributions of content co-occurrence or co-word literature reviews. International Journal of Consumer Studies, 48(2), e13031. https://doi.org/ 10.1111/ijcs.13031

[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] Opdahl, A. L., Al-Moslmi, T., Dang-Nguyen, D. T., Gallofré Ocaña, M., Tessem, B., & Veres, C. (2022). Semantic knowledge graphs for the news: A review. ACM Computing Surveys, 55(7), 1-38. https://doi.org/ 10.1145/3543508

[25] Ghai, D., Saxena, S., Dhingra, G., & Tripathi, S. L. (2025). A comprehensive review on performance-based comparative analysis, categorization, classification and mapping of text extraction system techniques for images. Multimedia Tools and Applications, 84(5), 2327-2484. https://doi.org/ 10.1007/s11042-024-20257-0