

Article

Deep Pre-Training Transformers for Scientific Paper Representation

Jihong Wang ¹, Zhiguang Yang ^{2,3} and Zhanglin Cheng ^{2,*}¹ School of Computer, Guangdong University of Education, Guangzhou 510303, China² Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China³ Xiaohongshu Inc., Shanghai 200001, China

* Correspondence: zl.cheng@siat.ac.cn

Abstract: In the age of scholarly big data, efficiently navigating and analyzing the vast corpus of scientific literature is a significant challenge. This paper introduces a specialized pre-trained BERT-based language model, termed SPBERT, which enhances natural language processing tasks specifically tailored to the domain of scientific paper analysis. Our method employs a novel neural network embedding technique that leverages textual components, such as keywords, titles, abstracts, and full texts, to represent papers in a vector space. By integrating recent advancements in text representation and unsupervised feature aggregation, SPBERT offers a sophisticated approach to encode essential information implicitly, thereby enhancing paper classification and literature retrieval tasks. We applied our method to several real-world academic datasets, demonstrating notable improvements over existing methods. The findings suggest that SPBERT not only provides a more effective representation of scientific papers but also facilitates a deeper understanding of large-scale academic data, paving the way for more informed and accurate scholarly analysis.

Keywords: representation learning; transformer; BERT; SPBERT; network encoding; scholarly big data



Citation: Wang, J.; Yang, Z.; Cheng, Z. Deep Pre-Training Transformers for Scientific Paper Representation. *Electronics* **2024**, *13*, 2123. <https://doi.org/10.3390/electronics13112123>

Academic Editors: Katia Lida Kermanidis, Phivos Mylonas, Arkaitz Zubiaga and Manolis Maragoudakis

Received: 9 April 2024

Revised: 10 May 2024

Accepted: 28 May 2024

Published: 29 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the field of research, scholars are tasked with the extensive work of searching, reviewing, and analyzing a variety of scientific papers pertinent to their fields to identify prior related studies. Nevertheless, the exponential increase in the availability of scientific publications is leading to information overload [1] and increasing the complexity of tasks, such as assessing scientific impact, recommending relevant literature, and finding specific papers of interest [2]. Natural language processing (NLP) and artificial intelligence (AI) technologies need to be used to improve search algorithms so that they can more accurately understand query intent and provide more relevant results to enhance the relevance and accuracy of search results. In the field of computational analysis, innovative methods are being developed to address the challenges associated with the scholarly examination of literature. The efficiency of such methods depends on the accurate computational representation of academic papers, a topic that has attracted considerable interest in recent years.

Strategies for the representation of scientific papers currently fall into three distinct categories, which differ in the techniques they employ: content-based, network-based and hybrid approaches. Traditional strategies often rely on the textual content of papers, represented by histograms such as those found in bag-of-words (BoW) or N-gram models.

The advent of deep neural networks has marked a significant advance in the field of natural language processing (NLP), improving the robustness of paper representations through the use of deep neural networks and extensive pre-training on large datasets.

In addition, a plethora of alternative methods have been developed to probabilistically model papers by exploiting their inherent network structure [3], bringing new dimensions to the analysis and representation of scientific literature.

While existing methods predominantly use citation and authorship data to analyse scientific literature, they often overlook the rich textual content of the papers themselves.

Citation networks form organically, leading to graph-based analysis methods, such as co-citation and bibliographic coupling relationship analysis, which focus on citation features by considering the relationships between papers. These methods increase the significance and specificity of the resulting feature vectors. Hybrid approaches combine these graph-based techniques with content-based strategies, using methods such as graph convolution networks (GCN) for network metadata and textual analysis for content metadata.

In contrast, our work addresses the full range of textual metadata within a paper, including title, abstract, body, and keywords, going beyond the limited scope of previous research, which typically only considers titles and abstracts. Specifically, we develop SPBERT, a language model rooted in BERT [4] and fine-tuned for scientific literature, to better capture the nuances of academic texts. SPBERT has undergone specialized pre-training tasks tailored to the domain of scientific research.

SPBERT is then used to derive text vectors from the title, abstract, and body of papers, integrating keyword information to mitigate BERT's limitations with long texts. To synthesize these different textual embeddings into a cohesive paper vector, we introduce two innovative aggregation methods. Together, these advances contribute to a more nuanced and comprehensive representation of scientific papers.

Our approach is evaluated against traditional representations in the context of paper classification and semantic search. The experimental results show that our method effectively captures paper representations. The main contributions of our work are outlined below:

- We propose SPBERT, a specialized pre-trained BERT-based language model fine-tuned on academic datasets, to improve the performance of NLP tasks in the domain of scientific paper analysis.
- We introduce a novel scientific paper encoder that can learn vector representations of each textual content by combining the keyword information and then aggregating it into a mixed model.
- We perform experiments on real scientific datasets to evaluate our proposed method, as well as an ablation study to confirm the effectiveness of each module.

2. Related Work

The foundation of downstream scientific natural language processing (NLP) tasks, such as paper classification, recommendation [5], and visualization, lies in effective paper representation [6]. These tasks have been developed to address the complexities associated with scholarly data. Content-based vectorization stands as the most widely adopted method for representing scientific papers [1].

A typical paper consists of various elements, including the title, authors, abstract, keywords, body, figures, and citation network. Content-based methods primarily extract and utilize text data. Traditional content-based representations have largely centered on titles or abstracts, employing techniques like bag-of-words (BoW) or N-gram models [7], which, unfortunately, lead to issues of high dimensionality and often overlook other textual elements.

The advent of unsupervised feature learning methods, utilizing deep neural networks to learn embeddings in lower-dimensional latent spaces, has marked a significant development in this field. For textual data, neural network methods such as word2vec [8] have been introduced, providing dense word and text representations through algorithms like Skip-gram [8] and Glove [9]. The text learning components in methodologies like Paper2vec [7] and VOPRec [1] draw upon word2vec and Doc2vec [10]. Nevertheless, these models typically learn context-independent word representations, which are static, thereby limiting their ability to address the nuances of polysemy.

Recent efforts have expanded to include the representation of words with contextual information. The emergence of new contextualized word embeddings, as referenced in studies on BERT and ELMo [4,11], enables the creation of nuanced representations for words based on their specific use cases, enhancing performance across a spectrum of

downstream tasks, such as question answering [12], sentiment analysis [13], and named entity recognition [14].

ELMo [11], which relies on a bidirectional language model pre-trained on extensive text corpora, has markedly advanced the benchmarks for many NLP challenges [15]. Building on this, the introduction of BERT [4], which employs a self-attention architecture widely celebrated in numerous NLP applications, has established itself as a leading method for word embedding. BERT's innovative approach to pre-training deep bidirectional representations from unlabeled text has set a new industry standard.

Further specializing the application of BERT, BioBERT [16], which has been trained on PubMed abstracts and PMC full-text articles, outperforms the standard BERT model in several biomedical tasks, including named entity recognition and relation extraction. In a similar vein, SciBERT [17] demonstrates superior performance over the original BERT on NLP tasks across various scientific disciplines by training with a corpus of scientific literature. These advancements illustrate that BERT is adept at extracting semantic content, and targeted training within a specific domain can significantly enhance model performance on related downstream tasks.

DeBERTa, as discussed in He et al. [18], enhances masked language modeling by utilizing a decoder optimized for predicting masked tokens, a departure from relying solely on the encoder. This improvement aids in comprehending and predicting missing word contexts, thereby enhancing the accuracy of language model pre-training. The performance of DeBERTa surpasses that of other models in various natural language understanding tasks, such as GLUE, SQuAD, and RACE benchmarks.

The subsequent version, DeBERTaV3 [19], further refines DeBERTa by incorporating the Replacement Token Detection (RTD) training loss from ELECTRA, alongside DeBERTa's disentangled attention mechanism. This amalgamation enhances pre-training efficiency and delivers notable performance enhancements in tasks like MNLI and SQuAD v2.0.

Transformer-based models, as highlighted in Rinaldi [20], are revolutionizing the field of NLP. Transformers excel in addressing semantic, syntactic, and even stylistic tasks. The pivotal factor in their success lies in pre-training on extensive corpora with abundant resources. By leveraging vast knowledge integration, these models offer a deeper understanding and analysis, leading to optimal results.

Heterformer [21] introduces a novel approach for node representation in text-rich, heterogeneous networks using the Transformer architecture, aiming to effectively manage diverse data types and enhance classification and link prediction tasks. Quintuple [22] introduces a quintuple-based learning model for bipartite heterogeneous networks, focusing on sophisticated representation by capturing complex interactions between diverse node types, significantly enhancing tasks like classification and prediction within these networks. Multi [23] presents a novel multi-view learning approach for representation learning in heterogeneous networks, effectively integrating diverse data types and views to enhance node representation, which significantly benefits tasks like node classification and link prediction. Hinormer [24] leverages a graph transformer to perform node representation learning in heterogeneous information networks, enhancing the capture of structural and semantic node information, which improves performance on various network analysis tasks.

T5 [25] introduces a unified framework that converts text-based language questions into a text-to-text format. This approach explores transfer learning technology in NLP and achieves top results in various benchmarks like summarization, question answering, and text classification. LongT5 [26] extends the T5 model to efficiently process longer text sequences. LongT5 utilizes a combination of local window attention and global sparse attention mechanisms, enabling the model to focus on crucial parts of the text without the typical computational burden of handling long sequences in transformers. It is versatile and suitable for various NLP tasks involving lengthy documents, such as document summarization, text classification, and long context question answering. Sentence-T5 [27] is specifically optimized for tasks at the sentence level. It adapts the T5 framework to emphasize generating or understanding sentences, making it potentially

more effective for activities like sentence similarity, paraphrasing, or other tasks focusing on individual sentences rather than longer texts. The output is primarily at the sentence level, potentially ideal for tasks that require concise and precise text generation or transformation.

KG-BART [28] is a generative common sense reasoning method that combines knowledge graphs and BART models. It enhances the model's ability to include relevant and accurate information in generated text by incorporating structured knowledge from the knowledge graph into the pre-trained BART model. This method is particularly suitable for tasks that require common sense reasoning, and can effectively improve the quality and information richness of generated text.

The K-LM [29] paper proposes a method to improve language model performance through knowledge enhancement in the academic field. This method combines domain-specific knowledge graphs with pre-trained language models to enhance the model's ability to process academic texts. K-LM places special emphasis on incorporating precise academic concepts and relationships into the model training process, thereby improving the accuracy and depth of the model in generating and understanding academic content.

3. Proposed Method

As shown in Figure 1, we develop a model for representing scientific papers, employing both BM25 [30] and SPBERT. SPBERT is distinct from BERT in that it is specifically pre-trained on scientific literature and is further refined through various tasks suited to this domain. Notably, our model utilizes a keyword-attention encoding technique to effectively encapsulate the essence of the scientific papers.

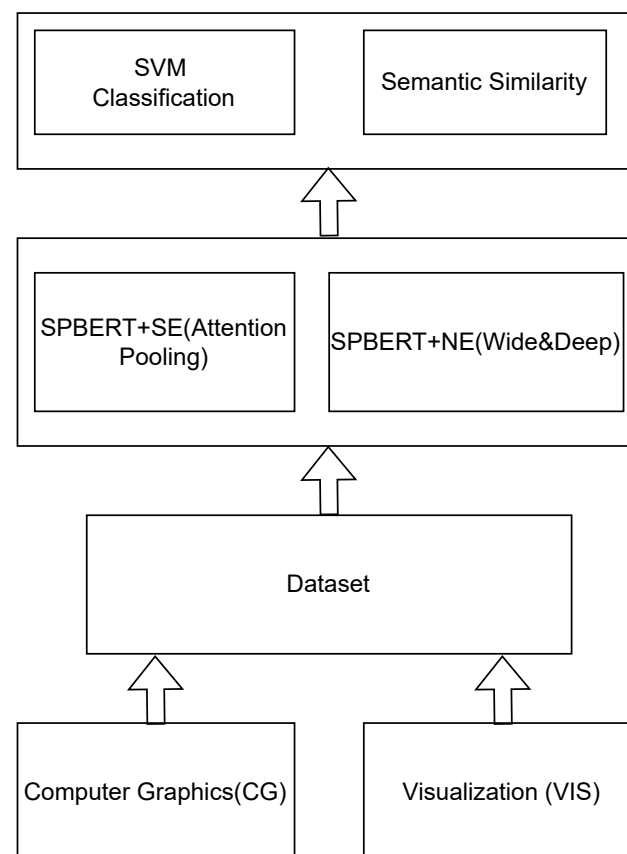


Figure 1. SPBERT model architecture.

As shown in Figure 2, the SPBERT model is derived from the BERT-Large model, and the model structure and parameters are basically the same (see the description of the Model section below). It uses computer graphics (CG) and visualization (VIS) datasets for self-supervised fine-tuning. It utilizes the class labels in these datasets to generate

SE (attention pooling) and NE (wide and deep) feature vector encodings, respectively. The model utilizes a support vector machine (SVM) approach for multi-class supervised learning. After training, the model is suitable for downstream tasks such as classification prediction and recommendation of similar items.

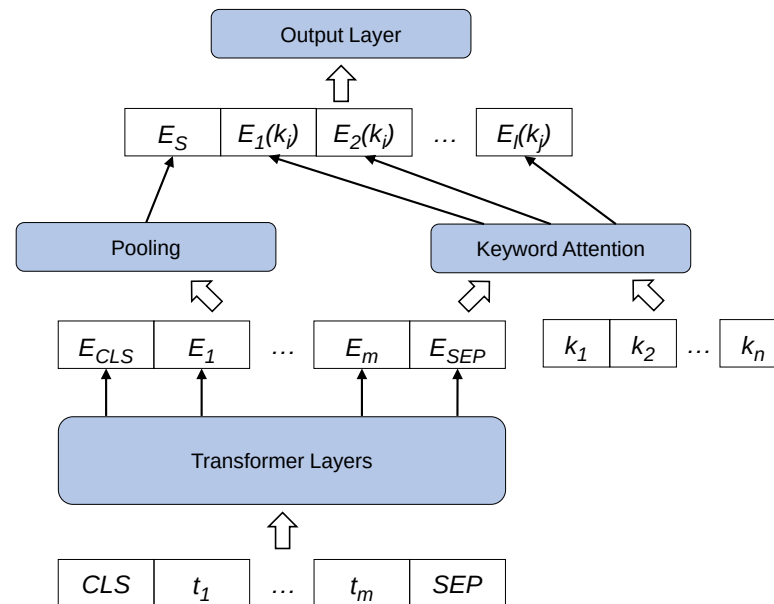


Figure 2. We use keyword attention pooling to combine the keywords with long text in document encoding.

3.1. SPBERT

3.1.1. Model

The architecture of BERT consists of a complex multi-layered bidirectional transformer encoder that utilizes self-attention mechanisms. The model is categorized into two types based on scale: the base and the large models. Our study employs the BERT_{Large} configuration, which is an English pre-trained variant with 24 layers, 768 hidden units per layer, 16 attention heads, and a total of 340 million parameters.

For the SPBERT model architecture, we specifically utilize the BERT-Large model. The model processes input as sequences of tokens, essentially fragments of contiguous text. Each sequence commences with a unique classification token ([CLS]). In instances where we have a pair of sentences, they are delineated using a distinct separator token ([SEP]).

BERT employs the WordPiece embedding [31] technique with a vocabulary size of 30,000 tokens to tokenize each element within the sequence in an unsupervised manner. We adopt the initial vocabulary provided with BERT, replacing the 995 ‘unused’ entries with the most frequent terms from our dataset used in the pre-training phase.

Beyond word embeddings, the input representation is further enriched with segment and position embeddings, which facilitate the model in understanding the structure and order of the input sequence.

3.1.2. Pre-Training Task

In the realm of learning representations for scientific papers, acquiring labeled data present a significant challenge. A principal benefit of models based on BERT is their ability to exploit a substantial corpus of unlabeled textual data for learning model parameters.

Our approach involves pre-training a neural network architecture that encompasses two key tasks: the masked language model (MLM) and the prediction of sentence relations (SRP). Task 1: Masked Language Model (MLM):

To cultivate a deep bidirectional representation, we randomly obscure certain input tokens. The model then forecasts these masked tokens, diverging from traditional left-to-right or right-to-left predictive training. Specifically, in BERT’s input layer, 15% of the

WordPiece tokens are randomly masked. For each hidden word, the model substitutes it (1) with the [MASK] token 80% of the time, (2) with a random token 10% of the time, and (3) leaves the original token unchanged 10% of the time. In the final layer, the model uses the masked words as a basis to predict the original tokens, applying cross-entropy loss. Through this technique, the model is trained to discern the contextual relationships among words. Task 2: Sentence Relation Prediction (SRP): Our approach diverges notably from BERT in the pre-training task. While BERT is engineered to perform general downstream tasks like question answering and semantic textual similarity, SPBERT is specifically calibrated for tasks associated with scientific papers. Moreover, recent studies have indicated that the Next Sentence Prediction (NSP) component, often deemed crucial, may actually have limited utility in many scenarios [32,33].

In the realm of paper representation, it is imperative to grasp the intricate relationships between sentences within a title or abstract. We model this as a four-category classification challenge. SPBERT generates sentence pairs as the basis for pre-training examples, upon which we predict the relational categories between these pairs. The categories are as follows: “0” indicates that the sentences originate from separate papers; “1” suggests that the sentences belong to the same paper, both from the abstract yet not consecutive; “2” denotes that the sentences are consecutive within the abstract of the same paper; and “3” implies that one sentence is the title while the other is from the abstract of the same paper.

This Sentence Relationship Prediction (SRP) task bears a resemblance to the Sentence Distance Task (SDT) employed in ERNIE 2.0. Through SRP, we are able to effectively mine information not only at the document level but also capture the nuances between titles and abstracts.

3.1.3. Pre-Training Data

We utilized SPBERT, previously pre-trained on a corpus of scientific papers from our earlier research, which includes references to VIS30K [34] and Vistory [35], among others. The selected papers predominantly cover topics within computer graphics, visualization, multimedia, and data mining. The papers have an average length of 197 sentences. For text preprocessing and sentence segmentation, we employed a modified version of ScispaCy.

3.2. Scientific Paper Encoder

Upon obtaining the textual encoding model SPBERT, we proceed to derive contextual representations from the textual data within scientific papers. Our approach emphasizes the importance of contextualized information while supplementing it with bag-of-words (BoW) data to bolster generalizability. Subsequently, these features are synthesized using a neural network, culminating in the generation of the definitive vector representation for each scientific paper.

Text data of the paper can be divided into two levels: words and sentences. Keywords always comprise n words $\{k_1, k_2, \dots, k_n\}$, while title, abstract, and body are ordered lists of sentences $\{s_1, s_2, \dots, s_a\}$; the difference among them is the number of the sentences a . Many prior studies have not adequately emphasized keywords, possibly due to the challenge of integrating standalone words with sentence-level textual information. However, deep learning models like BERT may struggle to distill key information from lengthy texts owing to a lack of robust supervised signals, and, while BERT generates an embedding vector for each token, the methodology for leveraging these at the sentence and document level remains a subject of ongoing research. Drawing inspiration from the work of Miao et al. [36] and Reim et al. [37], our proposed encoding model places greater emphasis on keywords in an attempt to address these challenges.

Specifically, suppose the input of SPBERT is the token sequence of the paper $\{CLS, t_1, \dots, t_m, SEP\}$, which contains m tokens and two special tokens.

SPBERT will output their embedding for each token like $\{E_{CLS}, E_1, \dots, E_m, E_{SEP}\}$. At this time, we check the corresponding keywords of the paper $\{k_1, k_2, \dots, k_n\}$. We first average pool all of the embedding to obtain E_S , which encodes raw information of the

token sequence without keywords attention, then focus on the l tokens which occur in the keywords bag to obtain their embedding $\{E_1(k_i), E_2(k_i), \dots, E_l(k_i)\}$, where $1 \leq i, j \leq n$, pooling them with E_s together to obtain the final encoding vector. With this “attention pooling”, we can combine keywords with the other textual data in a paper.

SPBERT serves as a pre-trained contextualized word embedding model, similar to its predecessors, ELMo [11] and BERT [4], when trained on extensive datasets. By utilizing these pre-trained weights, it achieves state-of-the-art results on various downstream tasks.

We use the SPBERT model described as Section 3.1 and “attention pooling” to obtain the contextual feature vector of the title F_s^T , abstract F_s^A , and body F_s^B . On the other hand, as a supplement to the contextual feature, we obtain BoW feature vectors F_b^T, F_b^A, F_b^B using BM25 [30]. When we do not have sufficient labels to fine-tune the model, the final paper representation vector FF is calculated as follows:

$$F_s^A \oplus (\alpha F_s^A + (1 - \alpha) F_s^B) \oplus (\gamma F_b^T + \beta F_b^A + (1 - \gamma - \beta) F_b^B) \quad (1)$$

where $\oplus$ is a concatenation operator, F_s^B can be seen as a supplement of F_s^A , and three BoW features share a common IDF vector so that they can be a weighted average.

As shown in Figure 3, our second method of encoding, referred to as Network Encoding (NE), is inspired by scenarios where an adequate number of labels are available, following the approach established by Cheng et al. The Wide and Deep model [38], as developed by Cheng et al. [38], integrates a wide single-layer component with a multilayered deep component. The wide component is crucial for the model’s ability to identify significant features, while the deep component contributes to the model’s generalizability. In addition to the previously mentioned features, we merge the title and abstract to form a composite feature “title+abstract”, which serves as an additional input, denoted as F_s^{T+A} , for the wide component of the model. The deep component of our model receives both contextual and bag-of-words (BoW) features, incorporating several hidden layers that are fully connected and employ rectified linear units (ReLU). This structure is particularly advantageous for exploring generalization capabilities.

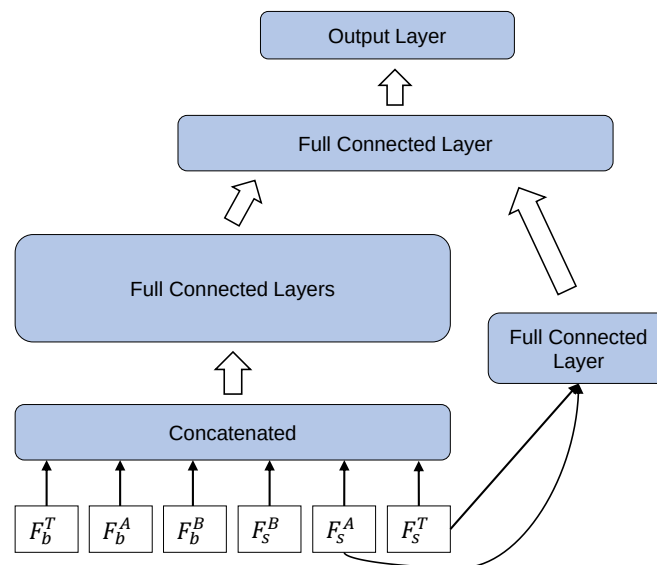


Figure 3. Network encoding model structure in paper features aggregation.

4. Experiment

4.1. Dataset

In our study, we analyze a corpus of 4379 research papers from the fields of computer graphics (CG) and visualization (VIS). These papers are sourced from two prominent conferences: 3289 from IEEE Visualization [39] and 1090 from SIGGRAPH [35]. Our analysis utilizes the titles, abstracts, main text, and keywords of these papers. For the

SIGGRAPH collection, we classify the papers into six sub-domain categories, such as animation/simulation, imaging/video, and modeling/geometry, etc., as delineated by Ke-Sen Huang. The VIS papers are grouped into three research track categories as follows: InfoVis, SciVis, and VAST.

4.2. Experimental Setup

Adhering to standard machine learning practices, we partitioned the papers into training, validation, and test sets in a 3:1:1 ratio. The optimization of the cross-entropy loss was performed using the Adam optimizer, combined with the implementation of a 0.5 dropout rate. Furthermore, we employed the BM25 [30] algorithm as implemented in the Gensim [40] library to select words for constructing the bag-of-words (BoW) feature vector, ensuring consistent dimensions across all papers.

The model experimental environment is a PC server, using the WIN10 operating system, the CPU is e5 2630v3, the memory is 32 g, and the graphics card is the nvidia m6000.

4.3. Evaluation Tasks

To assess and validate the performance of our learned embeddings, we focus on two pivotal tasks: categorizing papers and enhancing the precision of ranked information retrieval.

Paper Classification: Paper classification is a useful downstream task supporting users in finding and categorizing papers of interest. In this study, we address the task of categorizing scientific papers based on their vector representations, which essentially constitutes a multi-class classification challenge. Our methodology involves extracting features from the papers, followed by the application of a consistent classifier to predict their respective class labels. The effectiveness of our feature extraction technique is assessed by examining the classifier's performance in categorizing the papers. For this purpose, we employ Support Vector Machines (SVM) [41], a well-established model in the realm of supervised learning used for classification tasks. The choice of SVM, implemented using the scikit-learn library in Python, is due to our primary focus on the representation method rather than the classifier's intricacies. To ensure the reliability of our results, we present scores that represent the average across 10 distinct trials, thus minimizing experimental discrepancies.

Semantic Search: In the foundational application of text vector representation, assessing the efficacy of vector representations is crucial. Our methodology begins with feature extraction for each document. Subsequently, we convert query phrases into vectors using the same technique. The next step involves computing the normalized dot product between the query vector and each document's vector representation. By evaluating these dot products, we can gauge the similarity between vectors, thereby identifying the document most pertinent to the given query phrases.

In Equation (2), V_q is the vector of query, V_p is the vector of a paper, and $||V_p||$ is the norm of V_p .

$$\text{Similarity} = \frac{V_q * V_p}{||V_p||} \quad (2)$$

Furthermore, to assess the efficacy of our representation outcomes, we design a test set consisting of 100 search queries, formulated using standard terms from the field of computer graphics. For each query, we compute its similarity to every paper in our dataset, then rank the papers based on their similarity scores and select the top-K papers as candidates.

A paper is assigned a score of 1 if it is among these top-K candidates.

We detail the performance metrics at different levels, including Precision@1 and Precision@10, which reflect the accuracy of the top 1 and top 10 results as judged through manual evaluation.

To accurately gauge the system's capacity to present pertinent papers at the top of the candidate list, which users usually review first, we employ the mean reciprocal rank (MRR).

The MRR, defined in Equation (3), quantifies the average ranking position where the first relevant paper appears across all queries.

$$MRR = \frac{1}{n} \sum_{i=1}^n \frac{1}{rank(i)} \quad (3)$$

where $rank(i)$ is the highest-ranking where the first relevant paper i appears, and n represents the total number of target papers.

4.4. Baseline Methods

- **BM25** [30] is an improvement over simple BoW algorithms like TF-IDF [42] for representing documents in the vector space.
- **Doc2vec** [10] is proposed by Le et al. for learning latent representation for text documents, using a neural network. We compare this algorithm to SPBERT in obtaining pre-trained vectors.
- **BERT** is a recently proposed text representation algorithm and yielded state-of-the-art results over various tasks, including sentence classification and question answering. Therefore, we applied BERT as our strong baseline.

5. Results and Discussion

5.1. Quantitative Results

The summary statistics for test classification of various baselines, such as precision, recall, and F-measure, are presented in Table 1. Generally, BM25 [30] exhibits the poorest performance, likely attributable to its lack of consideration for contextual information. Comparatively, Doc2vec performs better than BM25 but falls short of BERT's performance, which stands out as the strongest baseline. Our proposed method surpasses all baselines, outperforming them by more than 10% in each metric. In conclusion, these classification outcomes indicate that our method demonstrates superior performance compared to the baseline models.

Table 1. Paper classification performance across models in each dataset.

Dataset	Method	Metrics		
		Precision	Recall	F1
CG	BM25 [30]	0.41	0.46	0.39
	Doc2vec [10]	0.46	0.53	0.45
	BERT [4]	0.60	0.67	0.62
	SPBERT+SE	0.69	0.74	0.71
	SPBERT+NE	0.72	0.78	0.74
VIS	BM25 [30]	0.48	0.51	0.47
	Doc2vec [10]	0.52	0.56	0.52
	BERT [4]	0.65	0.69	0.66
	SPBERT+SE	0.74	0.77	0.75
	SPBERT+NE	0.76	0.80	0.78

Table 2 displays the overall accuracy of semantic search results when compared to the baseline methods. As the retrieval list length Q increases, the precision of all techniques decreases—a common trend observed in most retrieval approaches. Doc2vec yields superior results compared to BM25, while BERT remains the most robust baseline.

Table 2. Paper semantic search performance across models.

Method	Metrics		
	Precision@1	Precision@10	MRR
BM25 [30]	0.35	0.32	0.24
Doc2vec [10]	0.43	0.40	0.29
BERT [4]	0.58	0.52	0.48
SPBERT+SE	0.72	0.65	0.64
SPBERT+NE	0.75	0.69	0.67

Our proposed model outperformed the BERT baseline according to the most widely used information retrieval metric. After integrating SPBERT and network encoding, the precision of all Q improved by over 15%, and the MRR increased by nearly 20%. This significant improvement serves as strong evidence that our approach can successfully establish the connection between research papers and associated queries at a semantic level.

5.2. Case Study

Table 3 displays two common instances of recommendations created by our algorithm. We exhibit the top three outcomes determined by the similarity of each document. In the case of “Real-time High-fidelity Facial Performance Capture”, all the suggested results align with the original paper’s focus on facial geometry and animation. In contrast, for the paper on liquid animation, two out of three recommendations are precise matches. The remaining suggestion introduces a liquid simulation technique that also employs the SPH method along with additional components.

Table 3. Top-3 recommendations generated for a query paper.

Title of the Query Paper	Top-3 Recommendation Results
Real-time High-fidelity Facial Performance Capture	1. High-quality Passive Facial Performance Capture using Anchor Frames 2. High-quality Single-shot Capture of Facial Geometry 3. Real-time Facial Animation with On-the-fly Correctives
Variational Stokes: A Unified Pressure-Viscosity Solver for Accurate Viscous Liquids	1. An Implicit Viscosity Formulation for SPH Fluids 2. Predictive-Corrective Incompressible SPH 3. Ghost SPH for Animating Water

6. Conclusions

In this paper, we have introduced SPBERT, a novel neural network-based framework designed for creating representations of academic papers. Our method can be easily applied to papers containing title, abstract, body, and keywords. To address the challenge of embedding lengthy documents and to capture keyword details effectively, SPBERT integrates keywords with other text information. Through a novel encoding network, which takes into account both contextual and bag-of-words (BoW) characteristics, we map the paper into a vector space. In empirical evaluations against three baseline methods, SPBERT demonstrated superior predictive performance in tasks such as paper classification and semantic search.

This method can not only be used in fields such as paper semantic representation, paper classification, and paper recommendation, but can also be used in other application scenarios, such as long text information extraction, summary generation, and semantic sorting, such as for book recommendations, judicial case recommendations, news reviews, etc.

A potential avenue for future research involves integrating our approach with graph-based data from citation networks [5] present in papers. Additionally, there has been limited exploration of incorporating figure information. Given the rapid advancements in computer vision research [43], enhancing representations to interpret figures could yield further improvements. Our ultimate goal is to develop a comprehensive approach that

amalgamates all publication metadata, encompassing text, graph-based data, and figure information.

With the swift integration and extensive use of multi-modal large models like GPT-4 [44], Llama2 [45], etc., it is now possible to extract additional information, such as images, based on text semantic representation. This enables the creation of more comprehensive feature representations, facilitating more accurate applications in paper classification, recommendation, and visual analysis.

Author Contributions: Methodology, Z.C.; Writing—original draft, Z.Y.; Writing—review & editing, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Guangdong Provincial Department of Education 2022 Higher Education Special Project grant number 2022GXJK287 and Shenzhen Science and Technology Program grant number GJHZ20210705141402008, APC was funded by Guangdong Provincial Department of Education 2022 Higher Education Special Project.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Author Zhiguang Yang was employed by the company Xiaohongshu Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Kong, X.; Mao, M.; Wang, W.; Liu, J.; Xu, B. VOPRec: Vector Representation Learning of Papers with Text Information and Structural Identity for Recommendation. *IEEE Trans. Emerg. Top. Comput.* **2021**, *9*, 226–237. [\[CrossRef\]](#)
- Xia, F.; Wang, W.; Bekele, T.M.; Liu, H. Big Scholarly Data: A Survey. *IEEE Trans. Big Data* **2017**, *3*, 18–35. [\[CrossRef\]](#)
- Nallapati, R.; Cohen, W.W. Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence of Blogs. In Proceedings of the ICWSM, Seattle, WA, USA, 30 March–2 April 2008.
- Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
- Jeong, C.; Jang, S.; Park, E.; Choi, S. A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics* **2020**, *124*, 1907–1922. [\[CrossRef\]](#)
- Nakazawa, R.; Itoh, T.; Saito, T. A Visualization of Research Papers Based on the Topics and Citation Network. In Proceedings of the 2015 19th International Conference on Information Visualisation, Barcelona, Spain, 22–24 July 2015; pp. 283–289.
- Ganguly, S.; Pudi, V. Paper2vec: Combining Graph and Text Information for Scientific Paper Representation. In Proceedings of the Advances in Information Retrieval, Aberdeen, UK, 8–13 April 2017; pp. 383–395.
- Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the 1st International Conference on Learning Representations, ICLR, Scottsdale, AZ, USA, 2–4 May 2013.
- Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
- Le, Q.; Mikolov, T. Distributed Representations of Sentences and Documents. In Proceedings of the 31st International Conference on Machine Learning, ICML, Beijing, China, 21–26 June 2014; Volume 32, pp. II–1188–II–1196.
- Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 1, pp. 2227–2237.
- Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 2383–2392.
- Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.; Potts, C. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1631–1642.
- Melamud, O.; Goldberger, J.; Dagan, I. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, Germany, 11–12 August 2016; pp. 51–61.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, U.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
- Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**, *36*, 1234–1240. [\[CrossRef\]](#)

17. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3615–3620.
18. He, P.; Liu, X.; Gao, J.; Chen, W. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv* **2020**, arXiv:2006.03654.
19. He, P.; Gao, J.; Chen, W. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv* **2021**, arXiv:2111.09543.
20. Ranaldi, L.; Pucci, G. Knowing knowledge: Epistemological study of knowledge in transformers. *Appl. Sci.* **2023**, *13*, 677. [\[CrossRef\]](#)
21. Jin, B.; Zhang, Y.; Zhu, Q.; Han, J. Heterformer: Transformer-based deep node representation learning on heterogeneous text-rich networks. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; pp. 1020–1031.
22. Zhou, C.; Chen, H.; Zhang, J.; Li, Q.; Hu, D. Quintuple-based Representation Learning for Bipartite Heterogeneous Networks. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–19. [\[CrossRef\]](#)
23. Chen, L.; Li, Y.; Deng, X. Multi-view learning-based heterogeneous network representation learning. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 101855. [\[CrossRef\]](#)
24. Mao, Q.; Liu, Z.; Liu, C.; Sun, J. Hinormer: Representation learning on heterogeneous information networks with graph transformer. In Proceedings of the ACM Web Conference 2023, Austin, TX, USA, 30 April–4 May 2023; pp. 599–610.
25. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
26. Guo, M.; Ainslie, J.; Uthus, D.; Ontanon, S.; Ni, J.; Sung, Y.H.; Yang, Y. LongT5: Efficient text-to-text transformer for long sequences. *arXiv* **2021**, arXiv:2112.07916.
27. Ni, J.; Abrego, G.H.; Constant, N.; Ma, J.; Hall, K.B.; Cer, D.; Yang, Y. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv* **2021**, arXiv:2108.08877.
28. Liu, Y.; Wan, Y.; He, L.; Peng, H.; Philip, S.Y. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 6418–6425.
29. Kumar, V.; Recupero, D.R.; Helaoui, R.; Riboni, D. K-LM: Knowledge augmenting in Language Models within the Scholarly Domain. *IEEE Access* **2022**, *10*, 91802–91815. [\[CrossRef\]](#)
30. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [\[CrossRef\]](#)
31. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
32. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
33. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Tian, H.; Wu, H.; Wang, H. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In Proceedings of the Thirty-Fourth Conference on Artificial Intelligence, AAAI, New York, NY, USA, 7–12 February 2020; pp. 8968–8975.
34. Chen, J.; Ling, M.; Li, R.; Isenberg, P.; Isenberg, T.; Sedlmair, M.; Moller, T.; Laramée, R.S.; Shen, H.W.; Wunsche, K.; et al. VIS30K: A Collection of Figures and Tables from IEEE Visualization Conference Publications. *IEEE Trans. Vis. Comput. Graph.* **2021**, *27*, 3826–3833. [\[CrossRef\]](#)
35. Dong, A.; Zeng, W.; Chen, X.; Cheng, Z. ViStory: Interactive Storyboard for Exploring Visual Information in Scientific Publications. In Proceedings of the 12th International Symposium on Visual Information Communication and Interaction, Shanghai, China, 20–22 September 2019.
36. Miao, C.; Cao, Z.; Tam, Y.C. Keyword-Attentive Deep Semantic Matching. *arXiv* **2020**, arXiv:2003.11516.
37. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3982–3992.
38. Cheng, H.T.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; Ispir, M.; et al. Wide & Deep Learning for Recommender Systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, Boston, MA, USA, 15 September 2016; pp. 7–10.
39. Isenberg, P.; Heimerl, F.; Koch, S.; Isenberg, T.; Xu, P.; Stolper, C.D.; Sedlmair, M.; Chen, J.; Möller, T.; Stasko, J. Vispubdata.org: A Metadata Collection About IEEE Visualization (VIS) Publications. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 2199–2206. [\[CrossRef\]](#)
40. Řehůřek, R.; Sojka, P. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 22 May 2010; pp. 45–50.
41. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
42. Joachims, T. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, Nashville, TN, USA, 8–12 July 1997; pp. 143–151.

43. Jang, Y.K.; Cho, N.I. Generalized Product Quantization Network for Semi-Supervised Image Retrieval. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3417–3426.
44. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
45. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Llama 2: Early Adopters’ Utilization of Meta’s New Open-Source Pretrained Model. *Preprints* **2023**, 2023072142.. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.