INTRODUCTION

1. Significance of this doctoral thesis

The number of scientific articles being published today is increasing at an unprecedented rate, posing significant challenges for researchers, particularly young and inexperienced ones, in identifying relevant and high-quality materials to cite. In the context of information overload caused by the vast number of scientific publications released each year, automatic citation recommendation systems have the potential to alleviate this burden. These systems can provide appropriate suggestions, enabling researchers to effectively navigate the massive volume of information.

Current approaches to the citation recommendation problem still exhibit several limitations. The first limitation lies in the fact that recommendation models do not fully exploit the available information in scientific articles. One of the pioneering studies in this domain was conducted by Ebesu [10] and Färber [11], who proposed a flexible architecture based on an encoder-decoder mechanism, known as the Neural Citation Network (NCN). While this model achieved superior performance compared to contemporaneous approaches on datasets such as RefSeer and arXiv CS, it still has notable shortcomings, particularly in its failure to comprehensively integrate critical information from articles, such as titles, authors, publication years, and venues, into the model training process.

The second limitation pertains to the insufficient utilization of the latest advancements in deep learning by existing citation recommendation models. For example, dual-step recommendation models such as DualLCR [12] and DualLCR-design [13], introduced by Medić and Šnajder in 2020 and 2022, respectively, still rely on Bidirectional Long-Short Term Memory (BiLSTM) mechanisms [14]. Similarly, the BERT-GCN model, developed by Jeong and colleagues [15], does not yet incorporate state-of-the-art advancements in natural language processing or citation-link graph analysis for scientific articles.

The third limitation concerns the fact that current citation recommendation models primarily focus on citation context and the content of candidate articles [16][17], while inadequately leveraging article metadata, including author names, publication years, and venues. These factors play a crucial role in shaping citation trends, as researchers tend to prioritize citing well-known authors, recent publications, or articles published in leading journals or conferences within their research domains.

2. Objectives of the doctoral thesis

The objective of the dissertation is to apply the latest advancements in deep learning models to develop a completely new model or propose solutions to enhance the performance of advanced citation recommendation systems.

3. Research subjects and scope of the doctoral thesis

The dissertation focuses on studying and analyzing several aspects related to the citation recommendation problem, including:

- Advanced deep learning models currently applied to the citation recommendation problem.
- Improvements in deep learning models, notable advancements in natural language processing, and diverse data representation methods for scientific articles.
- Performance evaluation metrics and datasets commonly used in advanced citation recommendation models.

4. Research methods

Theoretical research: Focus on studying and analyzing existing results from state-of-the-art citation recommendation systems, evaluating their strengths and weaknesses, and

proposing improvements to enhance the performance and accuracy of recommendation results. This involves leveraging deep learning techniques and models while also examining performance metrics and widely used datasets in citation recommendation models.

Experimental research: Implement and deploy source codes on widely used datasets in an experimental environment to measure and evaluate the results obtained from the proposed approaches.

5. Contributions of the doctoral thesis

With the goal of improving the performance of modern citation recommendation models, the dissertation has made the following significant contributions::

- Content-based filtering approach: Propose solutions to enhance the performance of the Neural Citation Network (NCN) model [10][11] (published in CT1).
- Content-based filtering combined with collaborative filtering approach: Construct a new model named RHN-DualLCR, which includes performance improvement solutions for the dual citation recommendation model DualLCR, previously introduced by Medić and Šnajder [12][13] (published in CT2 and CT4).
- Content-based filtering combined with graph-based filtering approach: Introduce a new citation recommendation model named SciBERT-GraphSAGE, which combines two recent advancements in natural language processing for scientific articles (SciBERT [18]) and graph structure representation (GraphSAGE [19]) (published in CT3 and CT5).

6. Structure of the doctoral thesis

The dissertation includes an introduction and the following main content chapters: Chapter 1 provides an overview of related studies and analyzes the limitations of prior research results. Chapters 2, 3, and 4 focus on the main contributions of the dissertation, with each chapter presenting proposed methods to improve the performance of modern citation recommendation models. The conclusion summarizes the main contributions of the dissertation, suggests future research directions, and highlights issues of interest to the author. Finally, the dissertation includes a list of published works by the author and references.

Chapter 1. OVERVIEW OF THE RESEARCH PROBLEM

1.1. Introduction to the citation recommendation problem

The citation recommendation problem was first introduced by McNee et al in 2002 [1]. According to this study, the typical operation of a citation recommendation model is described as illustrated in Figure 1.1:

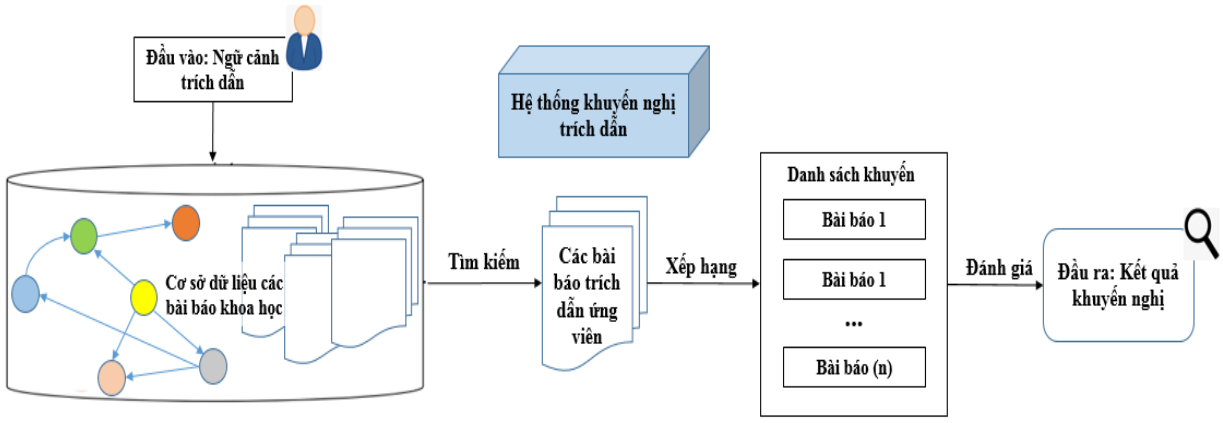

Figure 1.1. Flow diagram of a citation recommendation model

In general, the objective of a citation recommendation model is to suggest articles or citations to users by leveraging their preferences and research interests. Formally, a citation recommendation model can be defined as follows: (P) represents a set of articles that can be

recommended to researchers (U), and (Γ) is a utility function that measures the usefulness of an article (p_i) $\in$ (P) for a specific user (u_i) $\in$ (U). Mathematically, it can be expressed as (Γ) = (U) $\times$ (P) $\rightarrow$ (K), where (K) denotes the set of recommendations. For a user (u) $\in$ (U), the model suggests a subset of articles (p_i) $\in$ (P) that maximize (Γ) for that user, typically represented through rankings provided by the user.

1.2. Overview of related studies

Beel et al. [6] classified citation recommendation models based on the methods they employ into the following categories: collaborative filtering (*CF*), content-based filtering (*CB*), graph-based filtering (*GB*) and hybrid models.

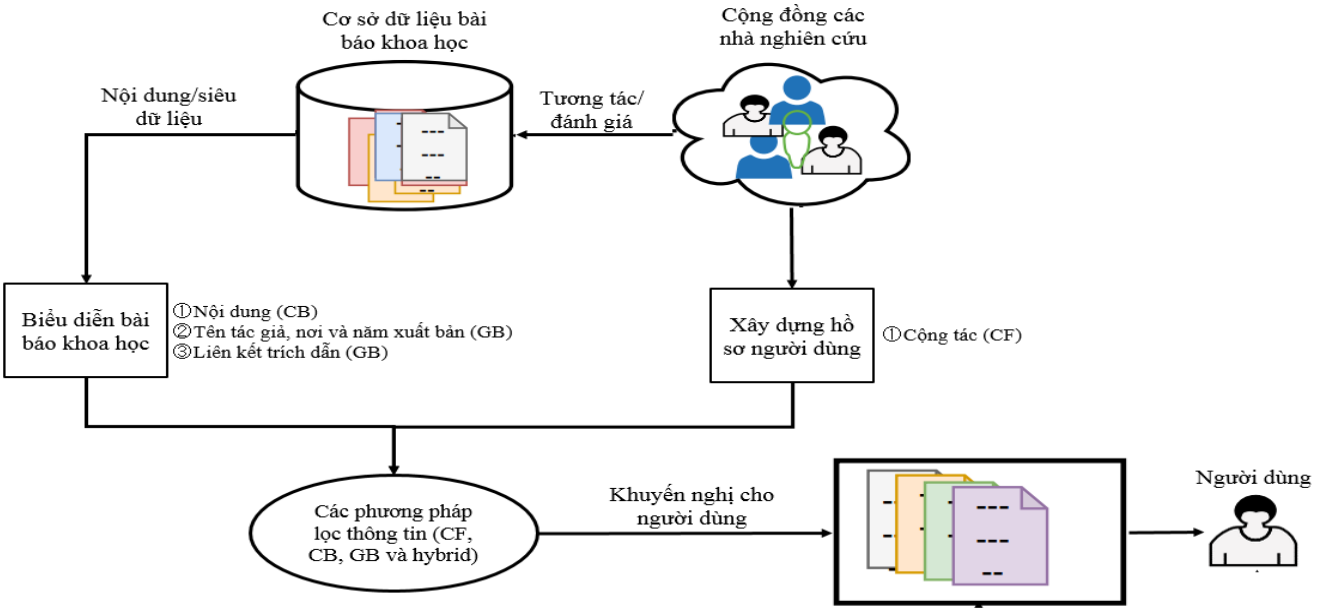

Figure 1.2. A citation recommendation model where article content and user profiles are exploited using various information filtering methods

1.2.1. Collaborative filtering (CF) models

Collaborative filtering models generate recommendations by leveraging users' past ratings along with the ratings from other users. The similarity between users and items is determined using a user-item rating matrix, which is maintained and updated regularly to ensure accurate recommendations. However, these models often face challenges with sparse data, particularly when there is insufficient rating information available for research documents [7][8][9].

1.2.2. Content-based filtering models

Content-based (CB) filtering models analyze the content of a query document to identify similar documents. This approach involves the following steps: ① Document embedding: transforming text into numerical vectors representing the content of the article (e.g., Doc2vec) ⇒ ② Nearest neighbor search: identifying the closest neighbors (potential citations) in the vector space ⇒ ③ Re-ranking potential citations: using ranking algorithms such as Okapi BM25 ⇒ ④ Recommendation: generating a ranked list of citations.

CB models focus entirely on the content of the article and do not rely on metadata such as publication venue, publication date, or citation count. This makes them particularly useful in cases where metadata is incomplete or unavailable [10][11][12][13][14]. However, these models have some limitations, such as not utilizing metadata, not fully incorporating the latest advancements in natural language processing, and not exploiting non-metadata information, such as the article title.

1.2.3. Graph-based filtering models

Graph-based filtering models utilize citation links to recommend relevant articles [15][16][17][18][19][20]. The process involves: ① Graph construction: creating nodes to represent articles and edges to represent citation links between them⇒②Node embedding: mapping articles into vector spaces using techniques such as GCN, HIN, GAT, or GraphSAGE... ⇒ ③ Similarity computation: calculating the similarity between embedded vectors to identify potential citations...⇒④Ranking: generating citation recommendations based on similarity scores.

This approach effectively exploits the relationships between articles through citation links, providing valuable insights into their relevance and impact within a research domain.

1.2.4. Hybrid models

Each type of model has its own strengths and weaknesses. Therefore, combining techniques from collaborative filtering (CF), content-based filtering (CB), and graph-based filtering (GB) has become an inevitable trend to maximize the extraction of information from articles. Representative studies following this approach include models such as DualLCR (CB+CF) [21][22], BERT-GCN (CB+GB) [23], MP-BERT4CR (CB+GB) [24], and RecCite (CB+CF) [25]. However, these hybrid models still have certain limitations, such as not fully utilizing supplementary information from articles or not exploiting the latest advancements in deep learning, particularly in natural language processing and graph convolutional networks.

Chapter 2. ENHANCED-NCN MODEL WITH ADDED TITLE INFORMATION AND BERT EMBEDDINGS

2.1. Introduction

Chapter 2 provides a detailed presentation of the proposed improvements to the NCN model developed by the research groups of Ebesu [10] and Färber [11]. These improvements involve incorporating additional article information and utilizing BERT embeddings. The results presented in this chapter have been published in CT1.

2.2. Analysis of limitations in the NCN model

The Neural Citation Network (*NCN*) is one of the first models introduced to address the citation recommendation problem. It was initially proposed in 2017 by Ebesu and Yi Fang [10], and later re-constructed in 2020 by Färber et al [11]. As illustrated in Figure 2.1, the NCN model consists of three main components: an encoder, a decoder, and an attention mechanism.

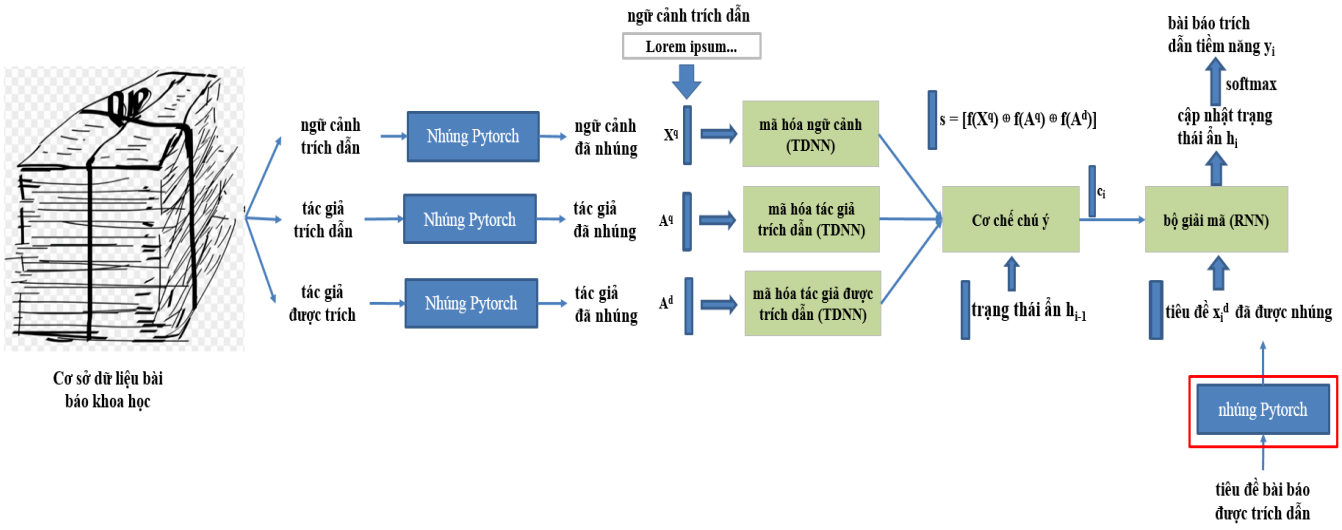

Figure 2.1. Overall architecture of the NCN model

2.2.1. Encoder

The encoder in the NCN model is designed to transform citation contexts and the names of cited or citing authors into representative features containing essential information about the corresponding context and authors. The encoder consists of two main components: citation context encoding and author encoding.

The citation context encoding component is responsible for encoding the citation context within scientific articles. This component utilizes a Time-Delay Neural Network (TDNN) introduced by the research group of Collobert [64]. TDNN enables parallel propagation through the network, allowing simultaneous computation of all feature maps. In the NCN model, TDNN comprises a convolutional layer, followed by a pooling layer and a fully connected layer.

To generate citation recommendations that include author information, the NCN model also integrates an author encoding component, which has a similar architecture to the citation context encoder. The author encoder is applied to (1) the embedding of the author names ($\boldsymbol{A}^q$) from the document in the query context and (2) the embedding of author names ($\boldsymbol{A}^d$) from all articles in the database. The author encoding process is performed iteratively using TDNN with varying receptive field sizes in the convolutional layer. The final representation of the text is denoted as the result of integrating the citation context encoding and author encoding components:

$$\mathbf{s}_j = [f(\boldsymbol{X}^q) \oplus f(\boldsymbol{A}^q) \oplus f(\boldsymbol{A}^d)]_j \tag{2.1}$$

where ($\boldsymbol{X}^q$) represents a citation context.

2.2.2. Decoder

The decoder in the NCN model is a Recurrent Neural Network (*RNN*) that utilizes Gated Recurrent Units (*GRU*) [65] as the gating mechanism, and it integrates an attention mechanism [66]. This decoder is applied to the titles of all potential documents that could be used as citations for a given query context. The primary function of the decoder is to generate scores for each document in the database, determining their relevance as citations for specific query contexts. These scores can then be used to recommend citations that are most suitable for the given query context.

2.2.3. Attention mechanism

The NCN model employs the attention mechanism initially introduced by Bahdanau et al. [66]. With this mechanism, encodings (s_j) generated by the citation context and author encoders are assigned weights based on the decoder's output (h_{i-1}) from the previous time step ($i-1$). The result is a context vector (c_i) computed as a weighted sum of the encoder outputs (s_j) according to their relevance. The attention mechanism is used to highlight particularly important encodings for the current time step. It is implemented as a Feedforward Neural Network (*FNN*) and culminates in a softmax layer, which transforms the attention vector (a_{ij}) into attention scores (α_{ij}). These scores indicate the importance of each encoder output (s_j) for the i-th word in the title of the document currently being decoded.

2.2.4. Limitations of the NCN Model

Although NCN is one of the most renowned citation recommendation models, having been cited in over 170 research works, it still has several significant limitations:

(1) Textual data embedding: The textual data of articles needs to be transformed into embeddings before being inputted into the encoder. However, the current NCN model uses the torch.nn.Embedding function from the PyTorch library for this transformation. While torch.nn.Embedding creates dense vector representations for discrete objects and is commonly used in natural language processing tasks to map categorical variables (like words or indices) to continuous vector spaces, this approach remains simplistic and does not leverage more advanced embedding techniques.

(2) Lack of article title integration: The article title is crucial as it encapsulates the most condensed meaning of the article's content. However, as illustrated in Figure 2.1, the current NCN model architecture does not integrate article titles into the encoding process. This limitation significantly impacts the model's performance, reducing its capacity to fully exploit the critical information embedded in article titles.

2.3. Enhancements to the NCN model

Based on the analysis of the limitations of the current NCN model, in this doctoral thesis implemented two enhancements to improve the model's performance: (1) Replacing the torch.nn.Embedding with BERT embeddings [54], which represent a more advanced achievement in natural language processing, and (2) Incorporating the titles of cited articles into the model for encoding. The enhancements to the NCN model are highlighted in red in Figure 2.2 below. The improved NCN model is referred to as Enhanced-NCN.

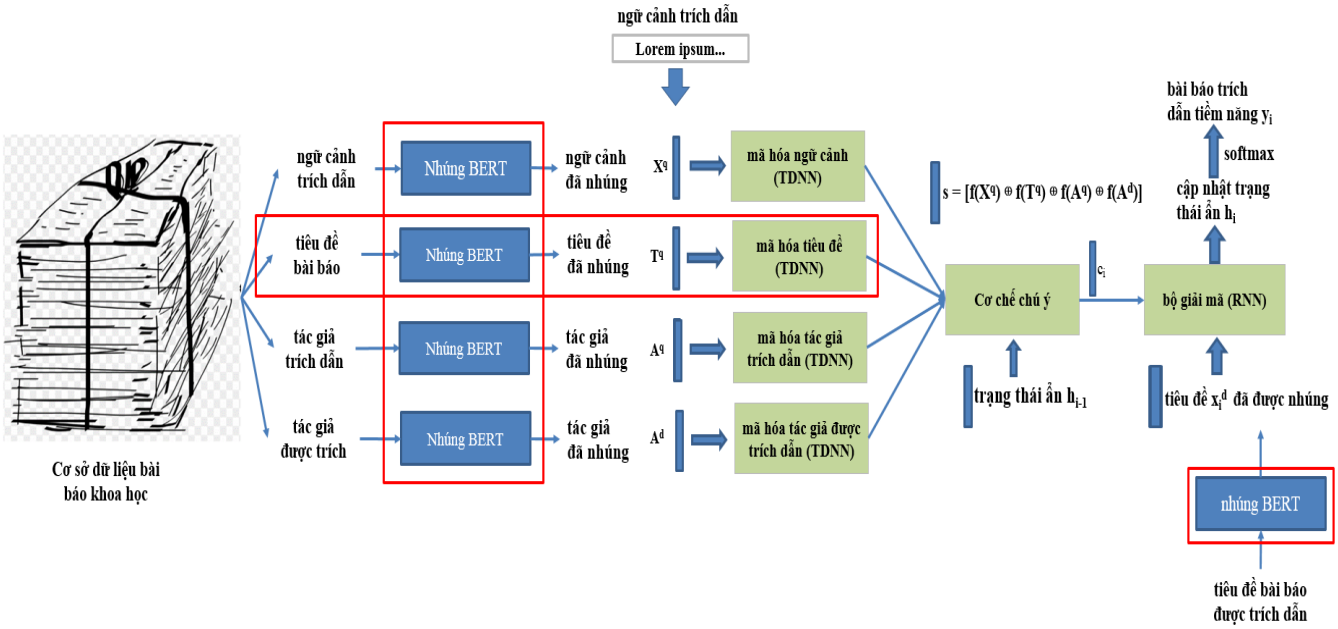

Figure 2.2. Overall architecture of the Enhanced-NCN model

2.3.1. BERT Embeddings

BERT (*Bidirectional Encoder Representations from Transformers*) is a machine learning technique based on the Transformer architecture, widely used for pretraining and natural language processing. BERT was introduced in 2019 by Jacob and colleagues at Google [54]. It is a powerful language model capable of generating context-sensitive embeddings for words and sentences derived from textual data. These embeddings are represented as low-dimensional vectors, which capture the meaning and relationships between words and sentences, thereby enhancing the performance of related tasks or models.

2.3.2. Incorporating article titles into the model

Although article titles are a critical factor providing relevant information for user queries and aiding recommendation systems in identifying suitable citation results, earlier versions of NCN developed by Ebesu and Yi Fang [10] and Färber [11] did not incorporate titles into the encoding process. Instead, these models focused on citation context, citing authors, and cited authors. To address this limitation and improve NCN's performance, this doctoral thesis integrated title encoding functionality into the Enhanced-NCN model. The final representation of the text processed by Enhanced-NCN combines citation context encoding, title encoding, and author encoding, as represented by:

$$\mathbf{s}_j = [f(\boldsymbol{X}^q) \oplus f(\boldsymbol{T}^q) \oplus f(\boldsymbol{A}^q) \oplus f(\boldsymbol{A}^d)]_j \quad (2.2)$$

where ($\boldsymbol{T}^q$) denotes the title of the article.

2.4. Experimental implementation of Enhanced-NCN model

2.4.1. Enhanced-NCN model construction

The Enhanced-NCN model was developed based on the source code of the NCN[1] model from the research by Färber et al. [11]. This was achieved by integrating BERT, one of the most advanced natural language processing tools, and adding a title encoder into the Enhanced-NCN architecture. This doctoral thesis implemented the model using Python 3.8.5 and PyTorch 1.7.1. For BERT, BertTokenizer and BertModel were utilized from the Python transformers library. Additionally, the Enhanced-NCN model employed the torchtext library[2] to convert datasets into formats suitable for PyTorch, facilitating preprocessing steps. Moreover, this enhancement leveraged the SpaCy[3] library in combination with torchtext for data tokenization and encoding.

After stemming and removing stopwords using SpaCy and the nltk[4] stopword set, the data was tokenized using BERT's vocabulary, which contains 30,522 tokens. This process was applied to the citation context, article titles, and citing/cited authors. To optimize batch processing, this chapter applied the bucketing technique, previously employed by Ebesu and Fang [10]. Similarly, in the decoding phase, the Okapi BM25[5] ranking function was retained to preselect titles for specific citation contexts, consistent with the original implementation of Ebesu and Fang [10].

2.4.2. Experimental dataset

Both the research work of Ebesu et al [10] and Färber et al [11] used two datasets, RefSeer and arXiv CS, in their studies. However, Färber et al. [11] noted that the RefSeer dataset could not be reconstructed from Ebesu et al.'s [10] research. Consequently, similar to Färber's approach, this dissertation evaluates the Enhanced-NCN model only on the arXiv CS dataset. The original arXiv CS dataset contains 1.7 million scientific articles from various subfields and topics within computer science. This dataset includes essential metadata for each article, such as titles, authors, abstracts, categories, and references. It supports multiple applications, such as trend analysis, citation recommendation, category prediction, knowledge graph construction, and semantic search. The dataset spans from January 1993 to April 2021 and is updated monthly. It is available in JSON format and can be downloaded from Hugging Face[6]. The reduced arXiv CS dataset, used for evaluating the Enhanced-NCN model, contains 502,355 records, including citation contexts, citing authors, article titles, and cited authors. The Enhanced-NCN model limited the length of citation contexts and article titles to 100 and 30 words, respectively, to balance model performance and training time. Unlike the study by Färber et al. [11], the Enhanced-NCN model incorporated article titles into the model for title encoding, significantly improving system performance. To train and evaluate the model, the arXiv CS dataset was split into 80% for training, 10% for validation, and 10% for testing.

2.4.3. Evaluation metrics

Most studies on citation recommendation problems utilize well-known evaluation metrics such as Normalized Discounted Cumulative Gain (*NDCG*), Mean Reciprocal Rank (*MRR*), Mean Average Precision (*MAP*), Recall@K, and Hits@K to assess model performance. In the study by Färber et al. [11], Recall@10 was used to evaluate NCN's performance. Accordingly, this doctoral thesis also employs Top@10 as the evaluation metric to measure the performance of the Enhanced-NCN model after improvements.

---

[1] https://github.com/timoklein/neural_citation
[2] https://pytorch.org/text/stable/index.html
[3] https://spacy.io/
[4] https://www.nltk.org/
[5] https://pypi.org/project/rank-bm25/
[6] https://huggingface.co/datasets/arxiv_dataset

2.5. Evaluation of experimental results with the Enhanced-NCN model

To determine the optimal parameters for the Enhanced-NCN model, this doctoral thesis adjusted four hyperparameters: data split ratio, number of layers, number of training epochs, and embedding size. These adjustments were made to compare the results with those obtained by the research group of Färber [11]. As shown in Table 2.1, this doctoral thesis systematically tuned the values of these parameters to identify the best configuration for the proposed Enhanced-NCN model. Citation contexts are the most information-rich components for generating citation recommendations. The titles of cited articles often contain critical details related to the cited content and are typically the first elements researchers notice when searching for references to cite in their work.

Table 2.1. Comparison of the results of the Färber et al 's NCN model [11] and the Enhanced-NCN model

Tên mô hình	Điều chỉnh	Split data	Number of layers	Epochs	Embedding size	Recall@10
NCN của nhóm Färber [11]	Embedding size	[0.8, 0.1, 0.1]	1	20	128	**0.0801**
		[0.8, 0.1, 0.1]	1	20	164	0.0663
		[0.8, 0.1, 0.1]	1	20	196	0.0527
		[0.8, 0.1, 0.1]	1	20	256	0.0413
	Number of layers	[0.8, 0.1, 0.1]	2	20	128	**0.1074**
		[0.8, 0.1, 0.1]	3	20	128	0.0867
Enhanced-NCN	Embedding size	[0.8, 0.1, 0.1]	1	20	128	0.0723
		[0.8, 0.1, 0.1]	1	20	164	**0.0921**
		[0.8, 0.1, 0.1]	1	20	196	0.0853
		[0.8, 0.1, 0.1]	1	20	256	0.0763
	Number of layers	[0.8, 0.1, 0.1]	2	20	164	**0.1285**
		[0.8, 0.1, 0.1]	3	20	164	0.1115

From the results in Table 2.1, it can be observed that when citation context information is preprocessed, and article titles are incorporated into the Enhanced-NCN model, the model gains access to richer information for generating recommendations. Consequently, this leads to significantly better performance compared to the results of Färber et al. [11]. Due to the increased input data volume in the Enhanced-NCN model, increasing the embedding size compared to the original NCN model yields favorable results. With the number of layers set to 1, this doctoral thesis experimented with embedding sizes of 128, 164, 196, and 256, achieving the best results with an embedding size of 164, where Recall@10 = 0.0921, compared to Recall@10 = 0.0801 (with an embedding size of 128) in Färber's model [11]. This result is illustrated in Figure 2.3.

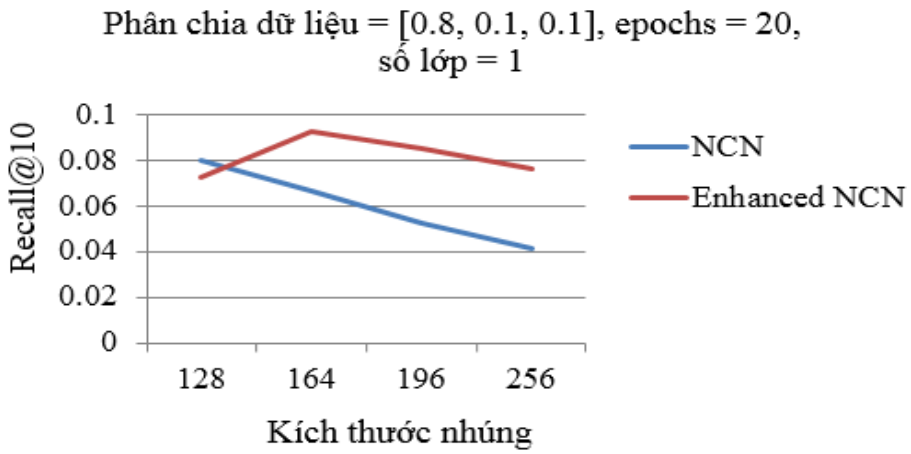

Figure 2.3. Comparison of Recall@10 results for the Enhanced-NCN model and Färber et al's NCN model [11] adjusted for embedding size

To optimize Recall@10 for the Enhanced-NCN model, this doctoral thesis further adjusted the number of layers in the model. Experimental results indicated that the best performance was achieved with 2 layers, applicable to both the NCN and Enhanced-NCN models. With this configuration, the Enhanced-NCN model achieved a Recall@10 of 0.1285, significantly higher than the Recall@10 of 0.1074 from Färber's NCN model [11]. However, increasing the number of layers to 3 or more did not yield further improvements. These results are clearly illustrated in Figure 2.4.

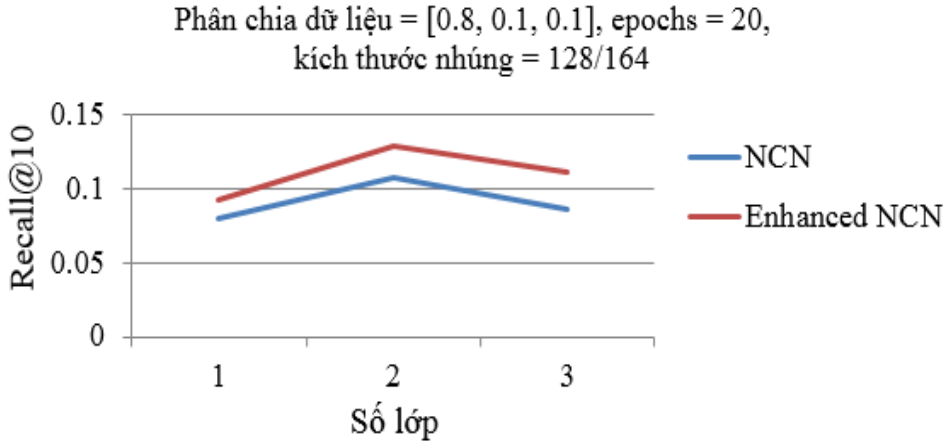

Figure 2.4. Comparison of Recall@10 results for the Enhanced-NCN model and Färber et al's NCN model [11] adjusted for the number of layers

2.6. Chapter 2 conclusion

In chapter 2, this doctoral thesis improved the Neural Citation Network (NCN) model, an advanced model initially introduced by the research group of Ebesu and Fang [10] in 2017 and later refined by Färber et al [11] in 2020. These enhancements included integrating the BERT model for preprocessing citation contexts and adding an encoder for article titles, enabling the titles to serve as a critical input source for the citation recommendation model.The performance of the Enhanced-NCN model was evaluated using the arXiv CS dataset. Experimental results demonstrated that the Enhanced-NCN model significantly outperformed Färber's NCN model [11] when evaluated using the same Recall@10 metric. Additionally, this chapter provided detailed analyses of how various parameters affected the performance of the Enhanced-NCN model. This information plays a vital role in optimizing the model and serves as a foundation for future research aimed at improving the efficiency of citation recommendation systems.

Chapter 3. RHN-DUALLCR MODEL WITH RECURRENT HIGHWAY NETWORK AND SCIBERT EMBEDDINGS

3.1. Introduction

Chapter 3 provides a detailed presentation of the RHN-DualLCR citation recommendation model, constructed by improving the proposed model of Medić and Šnajder [12]. The enhancements utilize Recurrent Highway Networks (RHN) and SciBERT embeddings to boost the model's performance. The findings in this chapter have been published in research works CT2 and CT4.

3.2. Analysis of limitations in the DualLCR model

The DualLCR model, introduced by Medić and Šnajder [12] in 2020, focuses on addressing the citation recommendation problem. While most previous methods relied solely on the text surrounding the citation location to represent the context [15][56][49][44][16], Medić and Šnajder proposed a contextual representation that integrates additional global information, such as the title and abstract of the cited article.

To generate citation recommendations for a specific context, the input to the DualLCR model consists of five types of information: (1) textual citation context, (2) the title and abstract of the article containing the citation (referred to as the citing article), (3) the title and abstract of the candidate article, (4) the list of authors of the cited article, and (5) the citation frequency of the candidate article over the past (y) years and its total citation count. The model's output is an overall recommendation score that indicates the suitability of the candidate article for citation within the given context.

The DualLCR model is structured into two main modules: the semantic module and the academic information module. The final recommendation score is a weighted sum of the scores generated by these two modules. The intuition behind the weighted sum is that, depending on the context, authors may prioritize citing influential papers within the research community (articles with high academic information scores) or papers that are directly relevant to specific details in their research (such as foundational theories or methods they are utilizing). Consequently, in the first scenario, the model assigns greater weight to the academic information score, while in the second scenario, it emphasizes the semantic score.

3.2.1. Semantic module

Similar to the work of Dai et al. [17], the DualLCR model employs Bidirectional Long-Short Term Memory (*BiLSTM*) [14] to represent the citation context, as well as to capture the content of the cited article and the global information from the cited article. Before being input into the semantic module, the text is segmented and tokenized using the SpaCy library. The target citation and other citations are masked with placeholders TARGETCIT and OTHERCIT, respectively. All three textual inputs are passed through two identical layers: a BiLSTM layer followed by an attention layer.

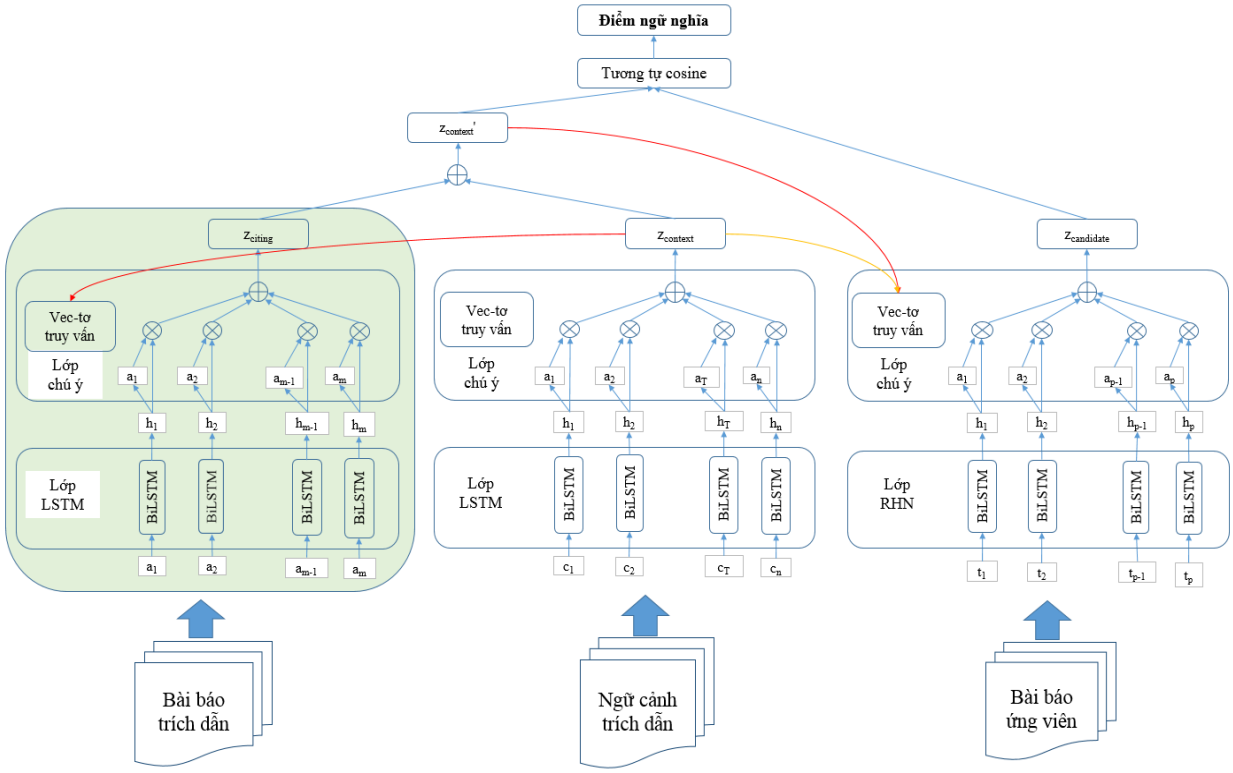

Figure 3.1. Structure of the semantic module in the DualLCR model [12]

Let (n) denote the total number of tokens in the input sequence, represented as $s = (t_1,..., t_n)$. Each token tit_iti is mapped to a d_e-dimensional embedding vector $x_i \in \mathbb{R}^{d_e}$ to form a sequence $x = (x_1,\ldots,x_n)$, using pre-trained embeddings from the embedding method of Bhagavatula [60]. The resulting sequence (x) is then passed through a BiLSTM layer with a hidden state size of (d_h), where the output ($\boldsymbol{h}_i$) at each step (i) is formed by concatenating the forward and backward hidden states: $\boldsymbol{h}_i = [\rightarrow \boldsymbol{h}_i; \leftarrow \boldsymbol{h}_i], \boldsymbol{h}_i \in \mathbb{R}^{2d_h}$. The hidden states of the input sequence sss are then passed through a subsidiary attention layer [66] to generate the final sequence embedding ($\boldsymbol{z}_s$). For an input query vector ($\boldsymbol{q}$) and a hidden state vector ($\boldsymbol{h}_i$) , the attention score for each step (i) is computed as:

$$a_i = \boldsymbol{v} \cdot tanh(W \cdot [\boldsymbol{q};\ \boldsymbol{h}_i]) \qquad (3.1)$$

Here, $\boldsymbol{v}$ and W are parameters of the DualLCR model. The normalized attention scores are applied to the corresponding hidden states and summed to produce the final sequence embedding ($\boldsymbol{z}_s$) . A different query vector (q) is used depending on the type of input. To make the contextual representation more specific to the citation being predicted, the DualLCR model utilizes the hidden state corresponding to the position of the citation's placeholder ($\boldsymbol{h}_T$) in the citation context. For the text of the citing paper, the DualLCR model uses the final sequence embedding of the context ($\mathbf{z}_{\text{context}}$), while for the text of the candidate paper being cited, the sum of the embeddings for the citing paper and the context ($\mathbf{z}_{\text{citing}} + \mathbf{z}_{\text{context}}$) is used. This enables the DualLCR model to focus on context-specific information within both the citing and cited papers. More specifically, by using the hidden state of the citation placeholder as the query vector to compute attention scores over the tokens in the citation context, the DualLCR model focuses on the relevant tokens in the context to obtain the transformed embedding of the citation position. Similarly, by using the embedding of the citation context as the query for the text of either the citing or cited paper, the DualLCR model focuses on the tokens in the text that are most relevant to the given citation context, as the text of a paper often describes different aspects of a research problem, not all of which are equally relevant to the current citation context.

Given the citation context (c) and the candidate paper (p), the semantic scoring function $s_{sem}(c, p)$ is defined as the cosine similarity between the enhanced context embedding and the transformed embedding of the candidate paper.

3.2.2. Bibliographic module

When the semantic context allows for multiple potential citations, Medić and Šnajder [12] suggested that authors generally tend to cite influential articles within the research community. This is the primary function of the academic information module. The module takes the author names and the citation count of the candidate article (p) as input and generates a single academic information score.

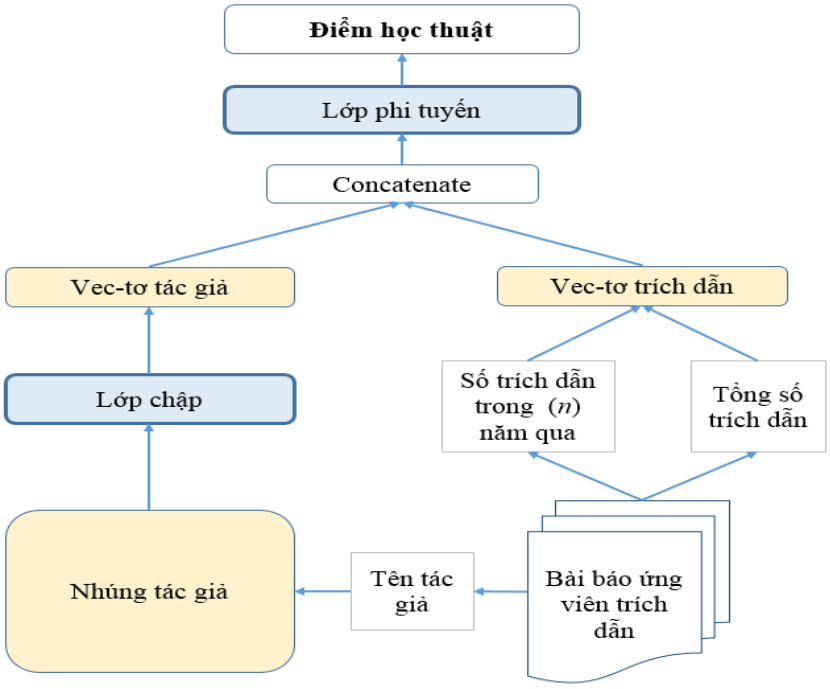

Figure 3.2. Structure of the bibliographic module in the DualLCR model [12]

The structure of the academic module is illustrated in Figure 3.2**.** Similar to the approach of Ebesu and Fang [10], the DualLCR model represents the names of the paper's authors in the form of embeddings. The list of author names for a given paper, denoted as $(a) = (a_1, \ldots, a_m)$, is first transformed into a sequence of author name embeddings, $(\boldsymbol{a_e}) = (a_{e1}, \ldots, a_{em})$. Next, the sequence $(\boldsymbol{a_e})$ is passed through a convolutional layer**,** followed by a max-pooling operation and a non-linear transformation to ultimately generate an embedding for the list of author names. This embedding is then combined with the total number of citations and the number of citations the paper has received in the past (y) years**.** Finally, the entire vector is passed through a non-linear layer to produce the academic information score $s_{bib}(p)$.

3.2.3. Final recommendation score

The final aggregated recommendation score $s_{fin}(c, p)$ is computed as the weighted sum of the scores $s_{sem}(c, p)$ and $s_{bib}(p)$. The weights for the scores are obtained by passing the transformed embedding of the citation context ($\mathbf{z}_{context}$) through a non-linear layer with two output values.

3.2.4. Limitations of the DualLCR model

While most previous citation recommendation models relied solely on the text surrounding the citation location to represent context, the DualLCR model [12] enhanced context representation by incorporating global information from the article. Specifically,

DualLCR included the title and abstract of the citing article in the context representation, which significantly improved performance compared to existing models. However, the DualLCR model still has the following limitations:

(1) Bidirectional Long-Short Term Memory (BiLSTM):

The DualLCR model still uses Bidirectional Long-Short Term Memory (BiLSTM) [14] to represent the enhanced context. Although BiLSTM is commonly employed in natural language processing tasks such as text classification and machine translation, it has several limitations:

- Gradient vanishing or exploding: BiLSTM can encounter issues with vanishing or exploding gradients during backpropagation, making the network difficult to train.
- High computational and memory costs: BiLSTM requires two LSTM layers for each direction and a concatenation layer to merge outputs, which can be computationally expensive and memory-intensive.
- Sensitivity to noise and outliers: BiLSTM assumes smooth and consistent input sequences, making it less effective when processing scientific texts, which are often noisy and inconsistent.
- Difficulty modeling long-term dependencies: Information from distant elements in the input sequence may become diluted or forgotten over time, limiting its ability to model long-term dependencies effectively.

(2) AI2 contextual embedding:

The DualLCR model uses embeddings from Bhagavatula's study [60] to encode the citation context and information (including the title and abstract) of both the citing and cited articles. These embeddings, known as AI2 embeddings, were introduced in 2017 by the Allen Institute for Artificial Intelligence (AI2). AI2 embeddings are character-based and can capture complex characteristics of word usage (such as syntax and semantics) and how these vary across linguistic contexts. Unlike traditional word embeddings, AI2 embeddings create representations that reflect the function of the entire input sentence, providing a richer understanding of word meanings. However, compared to SciBERT embeddings, AI2 embeddings exhibit several limitations:

- Context limitations: AI2 embeddings are trained on specific datasets, which may limit their effectiveness in domains or contexts not well-represented in the training data. For example, embeddings trained on general web text may perform poorly on specialized scientific texts.
- Static representation: If AI2 embeddings are of the traditional, non-contextual type (unlike BERT or SciBERT), each word is represented by a fixed embedding regardless of its context. This can be problematic for polysemous words, as the nuances of meaning in different contexts are lost.
- Bias: Pretrained embeddings may reflect and perpetuate biases present in the training data. This issue is critical when embeddings are used in decision-making processes or contexts where fairness and impartiality are essential.
- Out-of-vocabulary words: Pretrained embeddings struggle with unseen words not encountered during training, which is particularly problematic in fields like science and technology where new terms are constantly introduced.

3.3. Enhancements to the DualLCR model

Based on the analysis of the limitations in the current DualLCR model [12], this chapter applied two key enhancements to improve its performance: (1) replacing the Bidirectional

Long-Short Term Memory (BiLSTM) with Recurrent Highway Networks (RHN) and (2) replacing the AI2 embeddings introduced by Bhagavatula's study [60] with SciBERT embeddings [18]. The modifications to the DualLCR model are highlighted in red in Figure 3.3 below. The enhanced model is referred to as RHN-DualLCR.

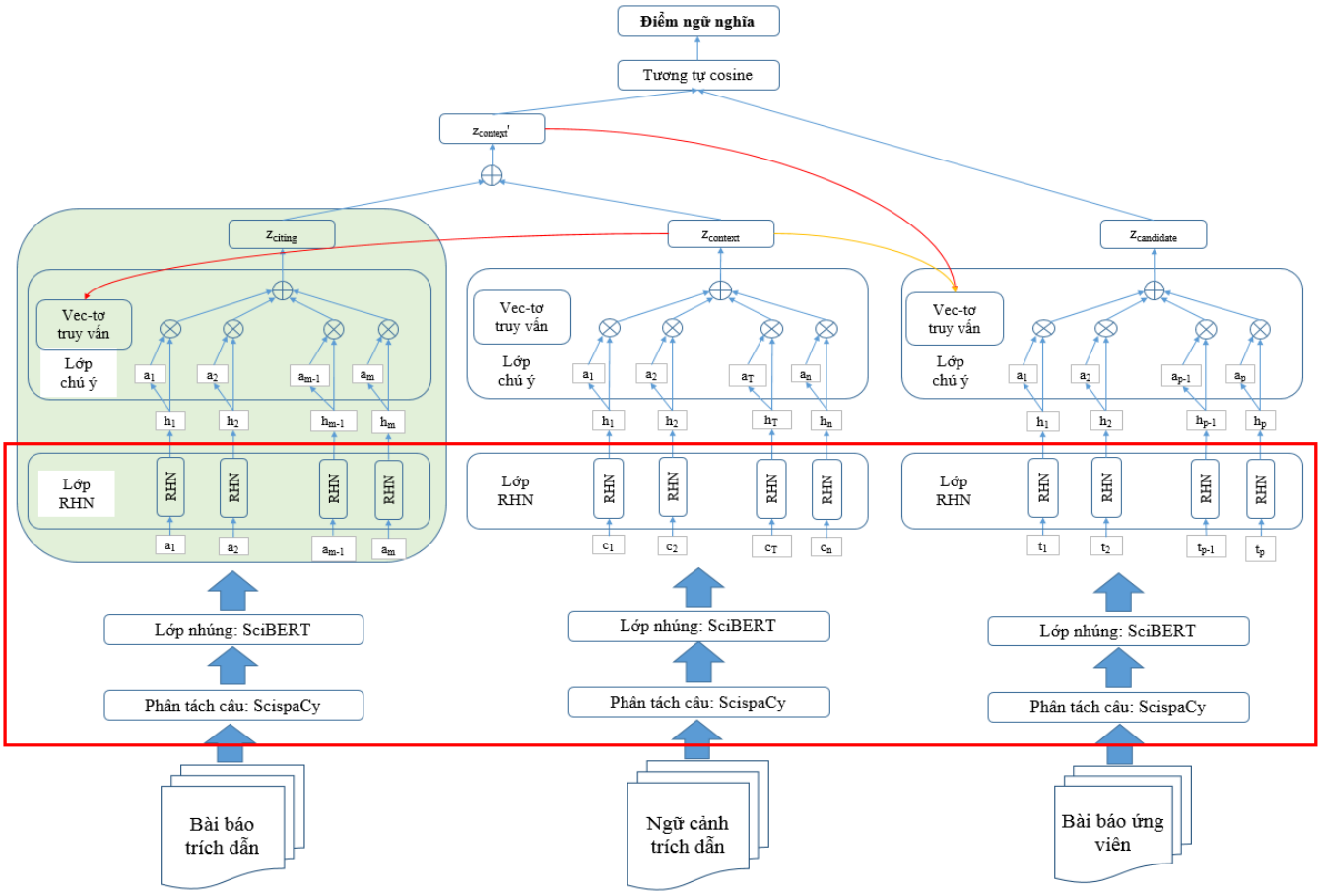

Figure 3.3. Structure of the semantic module in the RHN-DualLCR model

3.3.1. Recurrent highwaw network RHN

With training models, the nonlinear transition from one step to another in sequential processing tasks is highly complex, making the training of recurrent neural networks with "deep" transition functions challenging, even with Bidirectional Long-Short Term Memory (BiLSTM). To address this issue, Zilly et al. [20] introduced a theoretical study of recurrent networks using Geršgorin circle theory [68]. This study provided deeper insights into modeling and optimization while enhancing the performance of BiLSTM models. Based on this analysis, they proposed the Recurrent Highway Network (*RHN*), an extension of BiLSTM that enables deeper stepwise transitions. RHN is regarded as a robust model designed to leverage increasing depth in iterative transitions while maintaining the training simplicity of BiLSTM. The proposed architecture was evaluated through various language modeling experiments to demonstrate its performance and efficiency. For instance, in an experiment using the Penn Treebank corpus [69], increasing the transition depth from 1 to 10 reduced the word-level perplexity of the model from 90.6 to 65.4 while keeping the number of parameters constant. Furthermore, when evaluated on larger Wikipedia datasets [70] for character prediction tasks (text8 and enwik8), RHN outperformed all previous models, achieving an entropy of 1.27 bits per character. According to this doctoral thesis approach, applying RHN to citation recommendation models is expected to achieve more in-depth sequential contextual representations of relationships between target articles and candidate citations. Therefore,

implementing RHN is considered a promising direction for improving the effectiveness of existing citation recommendation systems.

3.3.2. SciBERT embeddings

SciBERT embeddings, introduced by Beltagy et al. [18], are a specialized variant of BERT designed specifically for scientific text. SciBERT was trained on a large corpus of scientific articles, research papers, and other academic content. Its primary goal is to capture the unique language and structures found in scientific literature, making it more effective for tasks involving the analysis of scientific texts. SciBERT has proven to be highly effective in natural language processing (NLP) tasks related to scientific texts, contributing to significant advances in extracting valuable information from scientific documents. SciBERT was trained on a corpus of 1.14 million research papers randomly selected from Semantic Scholar [26]. These papers include 18% in computer science and 82% from the broader biomedical field, with the full content of the articles (rather than just abstracts) used for training. The corpus averages 154 sentences per article, equivalent to 2,769 tokens, resulting in a total corpus size of 3.17 billion tokens. This size is comparable to the 3.3 billion tokens used to train BERT. Experiments conducted by the research team showed that for datasets like ACL-ARC and RefSeer, SciBERT significantly outperformed BERT [54]. The study presented in chapter 3 applied the SciBERT model to embed citation contexts, titles, and abstracts of both citing and cited articles before feeding them into the Recurrent Highway Network (RHN) within the semantic module. The remarkable performance of SciBERT in natural language processing tasks involving scientific texts is expected to enhance the effectiveness of the current DualLCR model for citation recommendation tasks.

3.4. Experimental implementation of the RHN-DualLCR model

3.4.1. Construction of RHN-DualCLR model

This doctoral thesis reconstructed the source code[7] of the citation recommendation system presented in the study by Medić and Šnajder [12], incorporating SciBERT [18], currently the most advanced model for processing scientific text, and replacing the existing BiLSTM layer with RHN [20]. The enhanced model was developed using Python version 3.8.5 and PyTorch version 1.7.1. For the pretrained SciBERT model, the new model utilized the AutoTokenizer and AutoModel modules from the Python Transformers library. Sentences within citation contexts and article abstracts were segmented using ScispaCy[8] [71], which has been optimized for scientific text processing. After separating, sentences from the articles were embedded using the SciBERT model.

3.4.2. Experimental dataset

To perform a comparison, the enhanced RHN-DualLCR model was evaluated on two datasets, RefSeer and ACL-ARC, as used in the original study by Medić and Šnajder [12]. Both datasets are widely utilized for measuring the performance of recently published citation recommendation systems [10][15][16].

Table 3.1. Statistics of the dataset by number of citation contexts and articles [12]

Dataset	Training	Validation	Test	Articles
ACL-ARC	30,390	9,381	9,585	19,711
RefSeer	3,521,582	124,551	126,021	624,957

[7] https://github.com/zoranmedic/DualLCR
[8] https://github.com/allenai/SciSpaCy

3.4.3. Evaluation metrics

In the study by Medić and Šnajder [12], two metrics are used to evaluate the performance of citation recommendation models: Recall Top@K and Mean Reciprocal Rank (MRR). Therefore, to compare the performance of the improved RHN-DualLCR model, this chapter also evaluates based on these two criteria.

3.5. Evaluation of experimental results

The research presented in this chapter focuses on enhancing the state-of-the-art citation recommendation model, DualLCR, proposed by Medić and Šnajder [12]. Accordingly, this section evaluates the experimental effectiveness of the improved RHN-DualLCR model in comparison to the DualLCR model. Similar to Medić and Šnajder [12], this section evaluates the citation recommendation performance using the standard metrics of Mean Reciprocal Rank (MRR) and Recall@K (R@K). According to the results reported by Medić and Šnajder [12], DualCon-ws achieved the best results in terms of MRR on the ACL-600 dataset. Furthermore, for the same dataset, DualEnh-ws obtained the best results for the R@10 metric. On the ACL-200 dataset, DualEnh-s achieved the best results for the R@10 metric, while DualEnh-ws achieved the highest MRR score. For the RefSeer dataset, DualEnh-ws performed best on both R@10 and MRR metrics. The experimental results of this study demonstrate that enriching citation context representations by incorporating global information is beneficial when the citation context is shorter. However, such enrichment is unnecessary for longer contexts, as longer contexts already provide sufficient information for citation recommendations.

Table 3.2. Comparison of results from Medić and Šnajder [12] and the enhanced RHN-DualLCR model.

Model		ACL-600		ACL-200		RefSeer	
		R@10	**MRR**	**R@10**	**MRR**	**R@10**	**MRR**
DualLCR [12]	**DualCon-ws**	0.689	**0.368**	0.647	0.335	0.406	0.206
	DualEnh-s	0.662	0.315	**0.716**	0.341	0.437	0.230
	DualEnh-ws	**0.699**	0.357	0.703	**0.366**	**0.534**	**0.280**
RHN-DualLCR	**DualCon-ws**	0.701	**0.391**	0.661	0.354	0.428	0.223
	DualEnh-s	0.683	0.342	**0.748**	0.363	0.461	0.256
	DualEnh-ws	**0.756**	0.379	0.718	**0.403**	**0.582**	**0.307**

Medić and Šnajder conducted an experimental study [13] to investigate the impact of three design choices in their previously proposed model [12]. Accordingly, this chapter also compares the results obtained from the RHN-DualLCR model with those from their experimental study to further highlight its contributions. Medić and Šnajder experimented with three design choices: the pre-filtering model parameters, training modes, and negative sampling strategies on two commonly used datasets, ACL-200 and RefSeer. Therefore, the RHN-DualLCR model is also compared with these design choices on the same datasets. For the ACL-200 and RefSeer datasets, RHN-DualLCR achieved the best performance with two variations: DualEnh-s and DualEnh-ws. As a result, this chapter focuses on comparing only these two variations. The comparison results are presented in Tables 3.3, 3.4, and 3.5, respectively.

Table 3.3. Comparison of performance of prefilter models BM25 and SPECTRE [13] with RHN-DualLCR

Mô hình		ACL-200		RefSeer	
		R@10	MRR	R@10	MRR
DualLCR-design [13]	**BM25**	0.254	0.077	0.173	0.055
	SPECTER	0.170	0.080	0.119	0.055
RHN-DualLCR	**DualEnh-s**	**0.748**	0.363	0.461	0.256
	DualEnh-ws	0.718	**0.403**	**0.582**	**0.307**

Table 3.4. Comparing the performance of Text+Bib reordering model for strict training regimes [13] with RHN-DualLCR

Mô hình		ACL-200		RefSeer	
		R@10	MRR	R@10	MRR
DualLCR-design [13]	**BM25(Text)**	0.531	0.268	0.300	0.116
	SPECTER(Text)	0.551	0.287	0.297	0.139
	BM25(Text+Bib)	0.725	0.401	0.324	0.147
	SPECTER(Text+Bib)	0.729	0.339	0.301	0.137
RHN-DualLCR	**DualEnh-s**	**0.748**	0.363	0.461	0.256
	DualEnh-ws	0.718	**0.403**	**0.582**	**0.307**

Table 3.5. Performance comparison of negative sampling strategies of DualLCR design [13] with RHN-DualLCR

Mô hình		ACL-200		RefSeer	
		R@10	MRR	R@10	MRR
DualLCR-design [13]	**Cited(Text+Bib)**	0.746	**0.413**	0.265	0.123
	Graph neighbors	0.676	0.363	0.418	0.210
RHN-DualLCR	**DualEnh-s**	**0.748**	0.363	0.461	0.256
	DualEnh-ws	0.718	0.403	**0.582**	**0.307**

To further demonstrate the performance of the RHN-DualLCR model, in addition to comparing it with the results of the two models, DualLCR [12] and DualLCR-design [13], published by Medić and Šnajder, this section also compares it with three state-of-the-art models for the citation recommendation task. These models are: (1) HAtten [16], which involves two stages: pre-retrieval and re-ranking; (2) NCN, a neural citation network proposed by Ebesu and Fang [10], along with its recent improvement by Färber et al [11]; (3) BERT-GCN [15], which combines BERT for text processing with the Graph Convolutional Network (GCN) [53] for encoding metadata of papers. The results of the model performance comparison are presented in Table 3.6.

Table 3.6. Comparing results from 3 state-of-the-art citation recommendation models with the RHN-DualLCR model

Mô hình	ACL-200		RefSeer	
	R@10	MRR	R@10	MRR
HAtten [16]	0.281	0.148	0.214	0.115
NCN [10] [11]	0.438	0.282	0.291	0.267
BERT-GCN [15]	0.685	0.378	0.423	0.281
RHN-DualLCR	**0.748**	**0.403**	**0.582**	**0.307**

3.6. Chapter 3 conclusion

In the research presented in this chapter, this doctoral thesis enhanced the existing citation recommendation model, DualLCR, originally published by Medić and Šnajder [12][13], by integrating recent advances in deep learning recurrent neural networks and achievements in natural language processing, particularly for the language used in scientific articles. This enhanced model incorporated ScispaCy [71] for sentence splitting in the citation context and article summaries, used the SciBERT [18] model to embed the scientific text of input papers, and replaced BiLSTM [14] with RHN [20], a model that extends the BiLSTM architecture to enable transitions across steps with greater depth.

The RHN-DualLCR model was evaluated the proposed model's performance on three datasets - ACL-200, ACL-600, and RefSeer - and achieved significant improvements compared to Medić and Šnajder's models presented in their original works [12][13], using the same evaluation metrics, R@10 and MRR. Moreover, the RHN-DualLCR model outperformed three advanced citation recommendation models: HAtten [16], NCN [10][11], and BERT-GCN [15]. Additionally, this chapter includes experimental tuning to examine how various hyperparameters affect the performance of RHN [20]. This chapter proposed ways to leverage these experimental results to further improve the model's effectiveness in the future.

Chapter 4. A NEW CITATION RECOMMENDATION MODEL USING SCIBERT AND GRAPHSAGE

4.1. Introduction

This chapter develops a novel context-aware citation recommendation model by combining two of the most advanced research achievements in representation learning for textual/contextual data and graph-based citation link representation: SciBERT [18] and GraphSAGE [19]. The findings presented in this chapter have been published in CT3 and CT5.

4.2. Challenges in current citation recommendation models

Most existing approaches to context-aware citation recommendation focus solely on the content of citation contexts and scientific papers [13][16][72]. This approach aims to bridge the semantic gap between these elements without considering information beyond the semantic content of scientific papers. However, scientific publications often include supplementary information, such as authors, conference/journal details, and publication years, which are critical for helping researchers understand the semantic similarities between scientific papers and citation contexts. For example: An author associated with a scientific paper may also co-author other related papers. Similarly, a conference or journal that publishes a specific scientific paper may also publish other papers on similar topics. Thus, incorporating information about authors and publication venues (conference or journal names) is expected to enhance the effectiveness of citation recommendation systems. Additionally, the publication year of a paper is also important for citation recommendation models. Specifically, researchers tend to cite the most recent and up-to-date papers when seeking citation materials. Based on these assumptions, this doctoral thesis found that the performance of context-aware citation recommendation is not solely influenced by the semantic similarity between citation contexts and scientific papers. It also depends on other factors, such as the authors, venues, and publication years of these papers.

Recent studies, such as Jeong et al.'s work [15], proposed the BERT-GCN model, which combines BERT for encoding citation contexts and paper content with a Graph Convolutional Network (GCN) for encoding metadata (authors, venues, and publication years) of papers. Other research efforts [11][12] have also begun to incorporate metadata into their models.

However, these models could still be improved by leveraging recent advancements in graph convolutional networks.

4.3. Construct a new citation recommendation model using SciBERT and GraphSAGE

This section details the development of a novel context-aware citation recommendation model by integrating two state-of-the-art research techniques for representation learning: SciBERT [18] for textual/contextual data and GraphSAGE [19] for graph-based citation link representation. As discussed in Chapter 3, SciBERT is a variant of the BERT [54] natural language processing model, specifically fine-tuned for tasks in the fields of scientific text and biomedical analysis. This chapter anticipates that using pre-trained SciBERT for contextual sentence representation will yield high effectiveness. Scientific data, such as research articles, often contain various metadata beyond textual content, including citation links between papers, authors, venues, and publication years. Therefore, this chapter applies the GraphSAGE model to represent citation links between papers and to derive learned embeddings from these relationships.

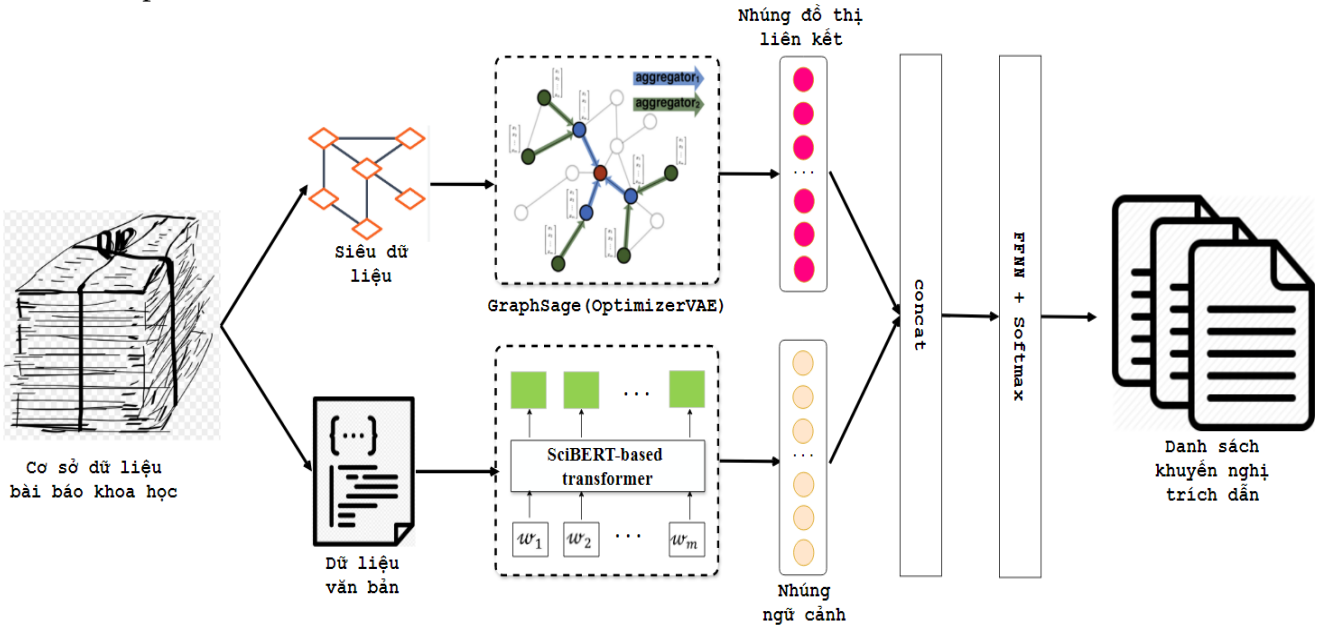

Figure 4.1. Overall architecture diagram of the SciBERT-GraphSAGE model

Figure 4.1 illustrates the architecture of the SciBERT-GraphSAGE citation recommendation model, which includes a context encoder for extracting text embeddings via SciBERT and a citation link encoder for generating graph embeddings using GraphSAGE. Both encoders are pre-trained using graph-based context and citation data from scientific articles. Subsequently, the data is input into these pre-trained models, and the embeddings generated by each encoder are combined. These combined embeddings are then passed through a feedforward neural network. The output is processed through a softmax layer, and the cross-entropy function is used as the loss function during the model training process.

4.4. Citation link graph encoder using GraphSAGE

The research of Hamilton et al [19] introduced GraphSAGE (***G**raph **SA**mpling and **AG**gregation*) as an extended and improved version of Graph Convolutional Networks (GCNs) [53]. The core concept behind GraphSAGE is to derive higher-order, inductive, and localized structural representations of nodes from the given citation graph. Unlike previous GCNs, GraphSAGE aggregates features for a target node based on a sampled subset of its neighboring nodes' attributes, rather than relying on the complete set of neighbors, as in GCNs [53]. This property makes GraphSAGE particularly suitable for citation recommendation applications.

Recent studies have demonstrated that GraphSAGE is a robust graph-based machine learning framework designed to efficiently learn representations from large-scale graphs. By representing academic articles as nodes in a graph, their metadata (authors, venue information, and publication years) as node attributes, and citation links between articles as edges, the citation neural network forms a natural graph structure where articles are interconnected through their citation relationships. Details of this process are depicted in Figure 4.2 below. Leveraging this inherent connectivity, GraphSAGE offers a promising approach to enhancing citation recommendations. It excels at learning node representations by sampling and aggregating information from neighboring nodes in the graph. In doing so, GraphSAGE not only captures the semantic significance of individual articles but also encodes the context provided by their citation relationships. As a result, the obtained node representations reflect the position of the articles within the broader academic landscape, capturing their similarities, influences, and thematic relationships.

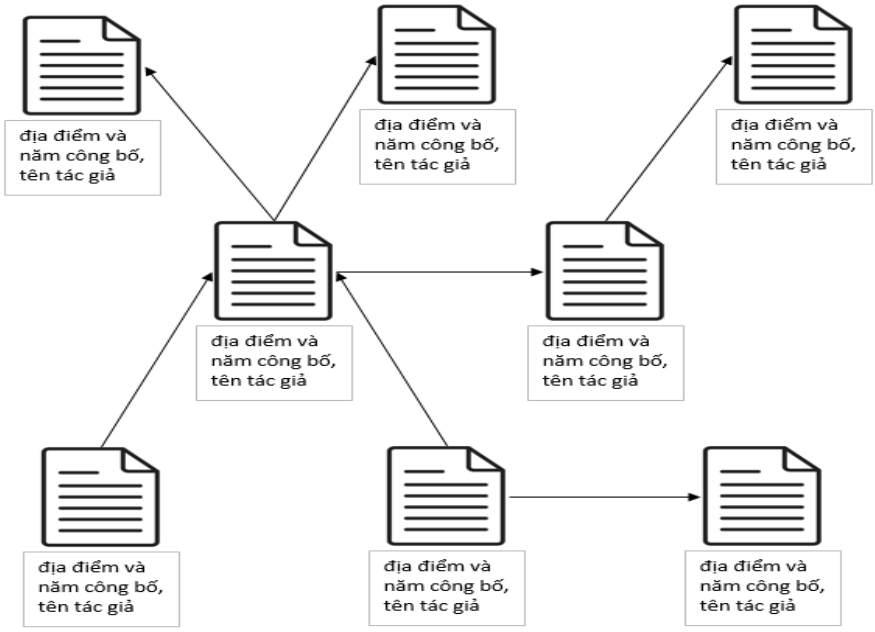

Figure 4.2. Generate nodes and edges for GraphSAGE from article metadata

4.5. Experimental implementation of the SciBERT-GraphSAGE model

4.5.1. Construction of the SciBERT-GraphSAGE model

This chapter developed the SciBERT-GraphSAGE model for the citation recommendation system by integrating the SciBERT citation context encoder with the GraphSAGE citation link encoder. To encode citation contexts, the candidate initialized SciBERT using the pre-trained model[9] provided by Beltagy et al. [18]. Similarly, the candidate modified the GraphSAGE source code[10] [19] to enable encoding of the citation link graph. All models were implemented using Python version 3.8.5 and TensorFlow version 2.7.0. Citation context embedding vectors and citation graph vectors were extracted using SciBERT and GraphSAGE layers, which were trained through separate learning processes. In the SciBERT model, the number of attention heads was set to 12, the encoder stack consisted of 12 layers, and the Adam optimizer [74] was used. The learning rate (η) was set to 0.0001, epsilon (ϵ) to 1, with beta parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a weight decay rate of 0.01. The model also had a maximum sequence length of 128, with zero-padding applied when the length was

[9] https://github.com/allenai/scibert
[10] https://github.com/williamleif/GraphSAGE

shorter than 128, and a hidden dimension size of 768. For the GraphSAGE model, the number of training epochs was set to 200, with the first hidden dimension corresponding to the number of papers in the dataset, and the second hidden dimension set to 768. The batch size was equal to the total document size (full-batch gradient descent was applied). The OptimizerVAE [73] was used as the optimizer, with a learning rate of 0.01.

4.5.2. Experimental datasets

This chapter evaluates the effectiveness of the novel SciBERT-GraphSAGE model on three standard datasets commonly used for local citation recommendation systems, namely: ACL-200 [12], RefSeer [12] and FullTextPeerRead [15]. These datasets were employed to assess the superior performance of SciBERT-GraphSAGE in comparison to five state-of-the-art citation recommendation models: CACR [72], BERT-GCN [15], HAtten [16], DualLCR [12] and DualLCR-design [13]. The statistical details of these three datasets are presented in Table 4.1.

Table 4.1. Statistics of 3 datasets (number of citation contexts and articles)

Datasets	Number of citation context			Number of articles	Publish year
	Training	Validation	Test		
ACL-200	30,390	9,381	9,585	19,711	2009 - 2015
FullTextPeerRead	9,363	1,043	6,841	4,898	2007 - 2017
RefSeer	3,521,582	124,551	126,021	624,957	- 2014

4.5.3. Evaluation Metrics

To assess the effectiveness of the SciBERT-GraphSAGE model, this chapter employs three commonly used evaluation metrics for citation recommendation systems: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and Recall Top@K.

4.6. Evaluation of experimental results

This section presents a comprehensive comparative analysis of the performance of the SciBERT-GraphSAGE model against several state-of-the-art local citation recommendation models that have recently demonstrated the best performance. The objective of this analysis is to rigorously assess the strengths and capabilities of the SciBERT-GraphSAGE model in comparison with existing advanced benchmarks. By conducting an in-depth comparison, the candidate aims to provide valuable insights into the performance improvements achieved through the proposed approach and highlight its potential contributions to the field. To demonstrate the superiority of the new model, the candidate subjected it to a rigorous benchmarking process against five widely recognized state-of-the-art models in the research community of local citation recommendation. The evaluation was conducted on three commonly used benchmark datasets: ACL-200 [12], RefSeer [12], and FullTextPeerRead [15]. These datasets encompass a variety of complex challenges, reflecting real-world scenarios and difficulties commonly encountered in local citation recommendation systems. By leveraging such diverse datasets, the candidate seeks to obtain a comprehensive understanding of the SciBERT-GraphSAGE model's performance across various contexts.

For the FullTextPeerRead [15] dataset, the SciBERT-GraphSAGE model was compared with the following advanced models: (1) CACR model [72], which integrates both a paper encoder and a citation context encoder based on the LSTM model; (2) BERT-GCN model [15], which combines BERT for text encoding and GCN for metadata encoding; (3) HAtten model [16], which incorporates a two-stage process consisting of a retrieval phase and a re-ranking phase; (4) SciBERT-base model [76], which is trained to predict the paper to be cited based on a given citation context. The comparative results are presented in Table 4.2 as follows:

Table 4.2. Performance comparison results of SciBERT-GraphSAGE model with 4 state-of-the-art models on FullTextPeerRead dataset

Mô hình	MAP	MRR	Recall@5	Recall@10	Recall@10	Recall@10	Recall@10
CACR [72]	0.1551	0.1549	0.2154	0.2761	0.4128	0.4794	0.5516
BERT-GCN [15]	0.4181	0.4179	0.4864	0.5291	0.6093	0.6495	0.6994
HAtten [16]	0.1672 *	0.1670	0.2780*	0.3060	0.4850*	0.5270*	0.5560*
SciBERT-base [76]	0.454	0.466	-	-	-	-	-
SciBERT-GraphSAGE	**0.5162**	**0.5163**	**0.6217**	**0.6744**	**0.7636**	**0.8099**	**0.8504**

In their published results, Gu et al. [33] did not present the evaluation results for their HAtten model using the MAP and Recall@K metrics, where (*K*) = 5, 30, 50, and 80. To ensure comparability, this chapter re-executed the experiments for the HAtten model[11] using the FullTextPeerRead dataset, and these newly obtained results are marked with an asterisk (*). Additionally, the authors of the SciBERT-base model [76] did not report their experimental results using the Recall@K metric. As shown in Table 4.3, among the recently published models, the experimental results for BERT-GCN [15] and SciBERT-base [76] on the FullTextPeerRead dataset yielded the most promising performance. However, by integrating SciBERT and GraphSAGE, which are more advanced and improved versions of BERT [54] and GCN [53], the SciBERT-GraphSAGE model demonstrated superior performance, surpassing these two models by 22% to 28% across all comparative metrics.

For the ACL-200 [12] and RefSeer [12] datasets, the SciBERT-GraphSAGE model was further compared against state-of-the-art research contributions recently published in the field of local citation recommendation: (1) HAtten model [16], which consists of a pre-retrieval phase and a re-ranking phase; (2) DualEnh and DualCon models [12], which leverage both semantic content and academic information data to assess the quality of each potential citation candidate; (3) DualLCR-design model [13], which has been optimized to achieve the best performance by fine-tuning its parameters. The comparative results are presented in Table 4.3 as follows.

Table 4.3. Performance comparison results of SciBERT-GraphSAGE model with 4 state-of-the-art models on 2 datasets ACL-200 and RefSeer

Bộ dữ liệu	ACL-200		RefSeer	
Mô hình	**MRR**	**Recall@10**	**MRR**	**Recall@10**
HAtten [16]	0.148	0.281	0.115	0.214
DualCon [12]	0.340	0.693	0.206	0.406
DualEnh [12]	0.366	0.716	0.280	0.534
DualLCR-design [13]	0.413	0.746	0.210	0.418
SciBERT-GraphSAGE	**0.472**	**0.765**	**0.308**	**0.565**

The experimental results presented in Table 4.3 indicate that, among recent publications, Medić and Šnajder [12] [13] have proposed models that achieve the best performance.

[11] https://github.com/nianlonggu/Local-Citation-Recommendation

Specifically, for the ACL-200 dataset, their DualEnh model [12] outperformed other models, while for the RefSeer dataset, the DualLCR-design model [13] demonstrated superior results. However, the SciBERT-GraphSAGE model achieved even better performance. Using the ACL-200 dataset, the SciBERT-GraphSAGE model outperformed previous models by 14% in MRR and 3% in Recall@10. Similarly, on the RefSeer dataset, the new model exhibited a 10% improvement in MRR and a 6% improvement in Recall@10, demonstrating its effectiveness.

4.7. Chapter 4 conclusion

This chapter presented a novel hybrid model for the local citation recommendation problem, named SciBERT-GraphSAGE, which integrates SciBERT and GraphSAGE. SciBERT [18] is a specialized version of BERT [54], specifically trained for tasks in scientific and technical research. Meanwhile, GraphSAGE [19] is a powerful graph-based machine learning framework designed for efficiently learning representations from large-scale graphs. By considering academic papers as nodes, metadata as node attributes, and citations as edges in the graph, the citation network forms a natural graph structure, where papers are interconnected through their citation relationships. Experimental evaluations were conducted using three widely used benchmark datasets (ACL-200 [12], RefSeer [12], and FullTextPeerRead [15]) against six state-of-the-art models (CACR [72], BERT-GCN [15], HAtten [16], DualEnh and DualCon [12], DualLCR-design [13], and SciBERT-base [76]). The SciBERT-GraphSAGE model consistently outperformed all competitors across three key evaluation metrics (MAP, MRR, and Recall@K). These results strongly validate the effectiveness of the proposed approach and confirm its significant contributions to the field of local citation recommendation.

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

1) Conclusion:

This doctoral thesis explores an approach that applies the latest advancements in deep learning models to the citation recommendation problem. The research has made the following key contributions

1. Proposed enhancements to improve the performance of the Neural Citation Network (NCN) model.
2. Introduced a new model, RHN-DualLCR, which incorporates multiple improvements to enhance the performance of the DualLCR two-step parallel network for citation recommendation, originally proposed by Medić and Šnajder [12][13].
3. Developed a novel citation recommendation model, SciBERT-GraphSAGE, by integrating two state-of-the-art techniques: SciBERT [18] for natural language processing and GraphSAGE [19] for citation link embedding using graph-based representations.

The findings from this research have been published in reputable specialized journals.

2) Future research directions:

For future research directions, the author of this doctoral thesis will focus on two primary areas:

1. Applying insights from heterogeneous networks and graph convolutional networks (GCNs)—developed by the research team—to the citation recommendation problem.
2. Conducting experiments on larger datasets, such as Microsoft Academic Graph[12], PubMed[13], and DBLP[14], to further validate and expand the applicability of the proposed models.

[12] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph
[13] https://www.nlm.nih.gov/databases/download/pubmed_medline.html
[14] https://dblp.uni-trier.de/xml/

LIST OF PUBLISHED ARTICLES RELATED TO THIS DOCTORAL THESIS

1. **Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, “Enhanced context-aware citation recommendation with auxiliary textual information based on an auto-encoding mechanism,” *Applied Intelligence*, 53(14), 2023, pp. 17381–17390, ISSN/eISSN:0924-669X/1573-7497, https://doi.org/10.1007/s10489-022-04423-1, Scopus indexed (Q2), SCIE IF: 5.086 (Q2).
2. **Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, "Enhancing local citation recommendation with recurrent highway networks and SciBERT-based embedding", *Expert Systems with Applications* (ESWA), Volume 243, 2024, 122911, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.122911 Scopus indexed (Q1), SCIE IF: 8.5 (Q1).
3. **Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Ngoc Thanh Nguyen and Bay Vo, "An effective context-aware citation recommendation model with SciBERT and GraphSAGE", *IEEE Transactions on Systems, Man and Cybernetics: Systems*, Volume 55, Issue 2, pp. 852-863, Feb. 2025, ISSN/eISSN: 2168-2216/2168-2232. https://doi.org/10.1109/TSMC.2024.3490774 Scopus indexed (Q1), SCIE IF: 8.7 (Q1).
4. **Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, "Enrich textual information for Hierarchical-Attention Text Encoder in Local Citation Recommendation”, *Kỷ yếu Hội thảo quốc gia lần thứ XXV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông* – Hà Nội, 8-9/12/2022 (National Symposium of Selected ICT Problems – VNICT(@) 2022), ISBN:978-604-67-2508-4.
5. **Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, "A context-aware citation recommendation model with SciBERT and GraphSAGE”, *Kỷ yếu Hội thảo quốc gia lần thứ XXVI: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông* – Bắc Ninh, 5-6/10/2023 (National Symposium of Selected ICT Problems – VNICT(@) 2023) ISBN: 978-604-67-2746-0.