Hierarchical Interdisciplinary Topic Detection Model for Research Proposal Classification

Meng Xiao, Ziyue Qiao, Yanjie Fu, Hao Dong, Yi Du, Pengyang Wang, Hui Xiong, Fellow, IEEE, and Yuanchun Zhou

Abstract—The peer merit review of research proposals has been the major mechanism to decide grant awards. However, research proposals have become increasingly interdisciplinary. It has been a longstanding challenge to assign interdisciplinary proposals to appropriate reviewers so proposals are fairly evaluated. One of the critical steps in reviewer assignment is to generate accurate interdisciplinary topic labels for proposal-reviewer matching. Existing systems mainly collect topic labels manually generated by principle investigators. However, such human-reported labels can be non-accurate, incomplete, labor intensive, and time costly. What role can AI play in developing a fair and precise proposal reviewer assignment system? In this study, we collaborate with the National Science Foundation of China to address the task of automated interdisciplinary topic path detection. For this purpose, we develop a deep Hierarchical Interdisciplinary Research Proposal Classification Network (HIRPCN). Specifically, we first propose a hierarchical transformer to extract the textual semantic information of proposals. We then design an interdisciplinary graph and leverage GNNs to learn representations of each discipline in order to extract interdisciplinary knowledge. After extracting the semantic and interdisciplinary knowledge, we design a level-wise prediction component to fuse the two types of knowledge representations and detect interdisciplinary topic paths for each proposal. We conduct extensive experiments and expert evaluations on three real-world datasets to demonstrate the effectiveness of our proposed model.

Index Terms—Natural language processing, classification algorithms.

I. INTRODUCTION

NOWADAYS, research funding is awarded based on the intellectual merits and broader impacts of proposals. Under this mechanism, investigators submit research proposals to open-court competitive programs managed by federal agencies (e.g., NSF, NIH). Proposals are then assigned to appropriate reviewers to solicit review comments and ratings. The biggest challenge of running such a peer-review system is to label precise interdisciplinary topics of proposals and allocate appropriate domain experts to obtain effective review comments. However, with the increasing number of universities, faculty members, hired graduate students, and publications, the submissions of research proposals have been exploding [1], [2]. For example, the National Natural Science Foundation of China (NSFC) organizes review panels for more than one million research proposals every year. Due to this growth momentum, there is an urgent need to introduce AI to automated and intelligent proposal labeling for review assignment.

In addition, research proposals increasingly exhibit interdisciplinary characteristics. It is inappropriate to select reviewers from only one field. It, therefore, is highly needed to classify a proposal into one or more disciplines, which we call the task of Interdisciplinary Research Proposal Classification (IRPC). Fig. 1 demonstrates an example of the IRPC task. After analyzing large-scale proposals from NSFC, we identify three unique data characteristics. These properties provide great potential to overcome the challenges in the IRPC task:

(Property 1: Hierarchical Discipline Structure) Before embarking on the specific problem, it is worth understanding how the funding administrator regiment the disciplines. In NSFC, thousands of disciplines and sub-disciplines are organized into a hierarchical structure to better index, search, and match reviewers and proposals. Each discipline is associated with a unique alphabetical identifier, which is called ApplyID. An ApplyID is a sequence of alphabetical letters. The ApplyID starts with a capital letter from A to H, representing eight major disciplines (e.g., Mathematical Sciences is marked by the letter A). The capital letter is followed by a sequence of numbers, in which every two numbers stand for a sub-discipline division. Fig. 1 (2) shows an example disciplinary path: F0601. F represents Information Sciences. 06 represents Artificial Intelligence that is a sub-discipline of Information Sciences. 01 represents

https://www.nsfc.gov.cn/publish/portal0/xmzn/2020/16/
**Fundamentals of Artificial Intelligence** that is a sub-discipline of artificial intelligence.

(Property 2: Interdisciplinarity) Interdisciplinary research is an indispensable part of advancing scientific progress [3]. In NSFC, an interdisciplinary research proposal will be marked by two ApplyID codes and assigned to the reviewers according to those codes. Fig. 1 (2) shows the ApplyID for an interdisciplinary proposal. This proposal is related to Fundamentals of Artificial Intelligence from the major discipline Information Sciences (F0601) and Neuroscience and Psychology (C09) from the major discipline Life Sciences. A proposal with incorrect or incomplete ApplyID codes will be assigned to wrong or non-expert reviewers, which lead to biased evaluations. Thus, to avoid the destruction of valuable pioneering and interdisciplinary research, it is critical to identify an accurate multiple disciplines for interdisciplinary proposals.

(Property 3: Multiple-type Text) The research proposal consists of multi-type textual data, including abstract, keywords, and title. Different types of text have distinct impacts on the topic path detection process. For example, an expert can identify the coarse-grained discipline (e.g., Information Sciences) of a proposal by reading the title. However, when the expert tries to identify a fine-grained disciplinary category (e.g., the Machine Learning, a sub-discipline of information sciences), they might need to consider the contents of each type (e.g., the introduction). In addition, since texts in most research proposals vary widely in length and structure, stitching each text together results in a severe loss of information.

To model the three unique proprieties, we propose three unique strategies for the IRPC task. First, we adopt the top-down Hierarchical Multi-label Classification (HMC) schema to utilize the hierarchical structure among the disciplines. This schema organize the prediction process level-wisely, thus reducing the candidate target number. Further, a level-wise prediction can generate the topic path under a given partial label, which is practical for a real-world application (e.g., the funding administrators or applicants might provide an unrefined topic path, and the model can generate the rest). Second, the interdisciplinarity among the disciplines can be represented as a weighted graph. We construct an Interdisciplinary Graph and use its topology structure to learn the label (discipline) embedding, thus incorporating the knowledge into each prediction step. Third, we model the semantic information within multiple-type text separately by a hierarchical Transformer architecture. In our design, each type of text will attach with a type token to preserve its type-specific semantic information. The self-attention mechanism in Transformer will also enhance the model’s ability to handle the texts with variate lengths. Based on the above technical insights, we propose **Hierarchical Interdisciplinary Research Proposal Classification Network (HIRPCN)**, a deep learning model with a hierarchical multi-label classification scheme.

In conclusion, our contributions are as follows:

- We identify the critical problems of assigning proposals to its related areas for review. We solve these problems via detecting topic paths on the hierarchical discipline structure.
- We propose HIRPCN, a framework integrating proposals’ semantic information and interdisciplinary topology knowledge for hierarchical label generation, which can generate one or more topic paths for non-interdisciplinary or interdisciplinary proposals.
- The experiments and experts’ evaluations show that our model can generate high-quality prediction results and reveal the candidate interdisciplines for research with incomplete labels.
- We post a demo\(^2\) of our model to demonstrate the process of the topic path detection and how it would aid the funding administrator. Our approach will apply to funding management in the future and contribute to improving the effectiveness and fairness of the science ecosystem.

II. **Definitions and Problem Statement**

A. Important Definitions

1) A Research Proposal: Applicants write proposals to apply for grants. Fig. 1 (1) shows a proposal which includes multiple documents such as **Title**, **Abstract**, and **Keywords**, **Research Fields**, etc. Let’s denote a proposal by \( A \), the documents in a

\(^2\)Demo available in http://101.35.54.56/
are two constant to weighting the two
\{E_pL \cup \Theta\} = \{\text{The only one greatest category}
\}\) is the proportion of topic overlap from
\(A\) discipline is a combination of
\(\in\) is the total number of the topic.
\([7]\) to mark the proper
\(h\)ierarchical
\(C\) and
\(\in\) based on the incomplete
\(K\)·
\(\{C_k\}\) is the set of the disciplines in
\(\text{the root level of } \gamma\). To describe the connection between
different disciplines, we introduce \(\prec\), a partial order representing
the \text{Belong-to relationship}, \(\prec\) is asymmetric, anti-reflexive and
transitive [4]:
- the only one greatest category \text{root} is the root of the \(\gamma\),
  \(\forall \gamma \in C_i, \gamma \in C_j, \gamma \prec \gamma_j \rightarrow \gamma_j \neq \gamma_i\),
- \(\forall \gamma \in C_i, \gamma \notin C_j, \gamma \prec \gamma_j\),
- \(\forall \gamma \in C_i, \gamma \in C_j, \gamma \in C_k, \gamma \prec \gamma_j \land \gamma \prec \gamma_j \rightarrow \gamma_i \prec \gamma_k\).

Finally, we define the Hierarchical Discipline Structure \(\gamma\) as a
partial order set \(\gamma = (C, \prec)\).

3) Topic Path: In this article, we define the target labels of each research proposal as the Topic Path \(L = \{L_0, L_1, L_2, \ldots, L_H\}\), \(L_i = \{l_i\}_{j=1}^{|l_i|}\) is a set of the labels in
the path labels on the \(i\)-th level (i.e., \(\forall l_i \in L_i \rightarrow l_k \in C_i\)).
\(H\) is the proper length of label paths. For example, in the
Fig. 1, the input research proposal are labeled by the ApplyID codes \(F00601\) and \(C09\). So the \(L\) for this research proposal can be processed as \(\{l_{\text{root}}\}, \{F, C\}, \{F06, C09\}, \{F00601, l_{\text{stop}}\}\).
The \(l_{\text{root}}\) stands for the root node of the discipline structure.
The \(l_{\text{stop}}\) is an \text{end-of-prediction} token, which is appended to the
last set in \(L\) to denote the topic path ended.

4) Interdisciplinary Graph: A discipline is a combination of
domain knowledge and topics. On the one hand, those topics will be
carefully selected [5] by the applicants to represent the key idea of their research proposal. On the other hand, the overlapping topic (co-topic) between disciplines could represent their similarity. So, how to measure interdisciplinary? S. W. Aboelela [6] believes that interdisciplinary research is engaging from two seemingly unrelated fields, which means: (1) these disciplines will have few overlapped topics. (2) those overlapped topics will be frequently cited by the researchers. In this article, we adopt the Rao-Stirling [7] (RS) to measure interdisciplinarity. The RS is a non-parametric quantitative heuristic, which is widely adopted to measure the diversity in science [7], [8], [9], the ecosystem [10] and energy security [11]. The RS is defined as:
\[
RS = \sum_{\alpha \neq \beta} (p_a \cdot p_b)^\alpha \cdot (d_{ab})^\beta,
\]
where the first part \(\sum_{\alpha \neq \beta} (p_a \cdot p_b)\) are proportional
representations of elements \(a\) and \(b\) in the system, and the rest
\(\sum_{\alpha \neq \beta} (p_a \cdot p_b)\) is the degree of difference attributed to elements \(a\) and \(b\). The \(\alpha\) and \(\beta\) are two constant to weighting the two components. We introduce the RS in a micro view to define the weight on \(e_{a \rightarrow b}\):
\[
e_{a \rightarrow b} = (p_{a \rightarrow b})^\alpha (d_{a \rightarrow b})^\beta,
\]
where \(p_{a \rightarrow b}\) is the proportion of the proposals in discipline \(a\) that contain same topics in \(b\), which represents the co-selected frequency of \(a\) to \(b\), \(d_{a \rightarrow b}\) is the proportion of topic overlap from \(a\) to \(b\). Their detailed definitions are as:
\[
p_{a \rightarrow b} = \frac{\sum_i \mathbb{1}(k_i \in \{K_a \cap K_b\}) F_a[k_i]}{\sum_i F_a[k_i]},
\]
\[
d_{a \rightarrow b} = 1 - \frac{|K_a \cap K_b|}{|K_a|},
\]
where the \(\mathbb{1}(\cdot) \rightarrow \{0, 1\}\) is the indicator function, \(K_a\) and \(K_b\) are topic set of \(a\) and \(b\), \(n\) is the total number of the topic. \(F_a\) and \(F_b\) are two lookup table holding the frequency of each topic being selected in the proposals that relat to \(a\) and \(b\).
By that, we build an Interdisciplinary Graph, denoted as \(G = (C, E)\), to represent the interdisciplinarity among disciplines. \(G\) is a collection of discipline set \(C\) and directed weighted edge set \(E = \{e_{a \rightarrow b}\}_{a,b=1}^{|C|}\). Each \(e_{a \rightarrow b}\) \(\geq 0\) represents the interdisciplinary strength from discipline \(a\) to discipline \(b\).

B. Problem Formulation

We formulate the IRPC task in a hierarchical classification schema and use a sequence of discipline-level-specific label sets \(L\) as topic path to represent the proposals’ disciplinary codes. To sum up, given the proposal’s document set \(D\) and the interdisciplinary graph \(G\), we decompose the prediction process into an top-down fashion from the beginning level to the certain level on the hierarchical discipline structure \(\gamma\). Suppose the \(k - 1\) ancestors labels in \(L\) are \(L_{\leq k} = \{L_0, L_1, \ldots, L_{k-1}\}\), where \(L_{\leq 1} = \{L_0\}\). The prediction on level \(k\) can be seen as a multi-label classification task on \(C_k\) based on the incomplete topic path \(L_{\leq k}\), this process can be defined as:
\[
\Omega(D, G, \gamma, L_{\leq k}; \Theta) \rightarrow L_k,
\]
where \(\Theta\) are the parameters of model \(\Omega\). Also, we use the end-of-prediction label \(l_{\text{stop}}\) in the last set \(L_{H}\) to mark the proper level (i.e., the \(H\)) where the model to stop on. Eventually, we can formulate the probability of the topic path assignment as:
\[
P(L_{\leq k} | D, G, \gamma; \Theta) = \prod_{k=1}^{H} P(L_k | D, G, \gamma, L_{\leq k}; \Theta)
\]
where \(P(L_{\leq k} | D, G, \gamma; \Theta)\) is the overall probability of the proposal \(A\) belonging to the label set sequence \(L_{\leq k}\), \(P(L_k | D, G, \gamma, L_{\leq k}; \Theta)\) is the label set assignment probability of \(A\) in level-k when given the previous ancestor \(L_{\leq k}\). In training, given all the ground truth labels, our goal is to maximize the (5).
III. INTERDISCIPLINARY TOPIC DETECTION

A. Overview of the Proposed Framework

Fig. 2 illustrate iteration of our framework upon four components: (1) Semantic Information Extractor (SIE) models the semantic components in research proposals according to their particular type (e.g., title, abstract, keywords, research fields) and encodes them into a matrix. (2) Interdisciplinary Knowledge Extractor (IKE) obtains the representation of previous prediction results from the pre-built interdisciplinary graphs. (3) Information Fusion (IF) fuses the type-specific semantic information and the representation of previous prediction results. (4) Level-wise Prediction (LP) predicts the probability that a research proposal falls within particular disciplines at the current step. After completing this iteration, our framework takes the discipline predictions, appends them to the label set sequence, and then moves to the next step. The initial input of the model will be the root node in the hierarchical discipline structure and a document set from a research proposal. The model will incrementally complete the input research proposal’s label path as the iteration goes on until the end-of-prediction label has been generated.

B. Semantic Information Extractor

Fig. 3 shows the overview of SIE. SIE comprises a Word to Vec and Multiple Positional Encoding (MPE) layer, and multi-layer SIEBlocks. Each SIEBlock consists of two parts: a word-level Transformer and a document-level Transformer. Formally, the representation of a proposal is obtained by:

$$D = N_c \times SIEBlock\left(\{W_i^{(0)}\}_{i=1}^N\right)$$

where the matrix $D$ is the representation of the input research proposal, $W_i^{(0)}$ is the word embedding matrix of the document $d_i$. The $N_c$ is the total layer number of SIEBlocks.

Our Perspective. We consider two essential aspects of the research proposal. First, different types of documents (e.g., title, abstract, keywords, research fields) in a research proposal should be separately weighed when identifying the topic path of the proposal. Second, in the proposal, each document varies in length (e.g., title and abstract), and concatenating them together will cause a severe loss of information. Inspired by the long-text modeling methods [13], [14], [15], we design a hierarchical Transformer architecture to embed both texts and their particular type within a proposal into a matrix.

Step 1: Word to Vec and Multiple Positional Encoding. We first vectorize each words in document by a pre-trained $h$-dimensional Word2Vec [16] model, then sum them with positional encoding to preserve the position information. After separately concatenating each document’s word vector, we can obtain the word embedding. It is worth noting that we padded each document to an identical length before the Word to Vec component. The length of each document after padding is denoted by $|d|$. The word embedding for document $d_i$ is denoted as $W_i^{(0)} \in \mathbb{R}^{|d| \times h}$. Then we separately feed each document representation matrix into the multi-layer SIEBlocks.

Step 2: Word-level Transformer. In the first stage of $l$-th SIEBlock, we aim to use a word-level Transformer to extract the contextual semantic information. We can obtain the document representation matrix $W_i^{(l-1)} \in \mathbb{R}^{|d| \times h}$ from the previous layer. Then, we can formulate the process of the current layer as:

$$W_i^{(l)} = Transformer_w(W_i^{(l-1)})$$

where the $W_i^{(l)} \in \mathbb{R}^{|d| \times h}$ is the vanilla document representation matrix in current layer. The $Transformer_w(\cdot)$ is the word-level Transformer block in $l$-th SIEBlock.

Step 3: Document-level Transformer. In the second stage of $l$-th SIEBlock, we aim to fuse each vanilla document representation matrix with its correlated type information by a document-level Transformer:

$$D_i^{(l)} = Transformer_d(\{d_i^{(l-1)} \odot W_i^{(l)}\}_{i=1}^N)$$

where $Transformer_d(\cdot)$ is the document-level Transformer block. $d_i^{(l-1)} \in \mathbb{R}^h$ is the type-specific representation of $d_i$ from previous layer. In the first layer, the model will initialize each document’s type-specific representation with a randomly initialized look-up table according to their type $t$. And the $\odot$ is a fusing operation.
Inspired by ViT [17] and TNT [18], we set the \( \odot \) operation as: (1) a Vectorization Operation on \( W^{(l)} \), (2) a Fully-connected Layer to transform the vanilla document representation from dimension \( |d|h \) to dimension \( h \). (3) an Element-wise Add with the vector \( d^{(l-1)} \) from previous layer to obtain the type-specific representation in current layer. \( D^{(l)} \in \mathbb{R}^{|T| \times h} \) is the outputs of current layer, which can be seen as \( |T| \)-views of high-level abstraction for proposals. After \( N_c \) times propagation, we can obtain the final output of SIE as \( D = D^{(N_c)} \).

\[ H_i = N_g \times GCN(E_i, H^{(0)}_i), \]  
\[ e_i = \text{Readout}(H_i, L_i), \]

where the \( \text{Readout} \) layer is a lookup operation (i.e., take the \( L_i \) included node embedding) followed by a mean pooling layer. The \( e_i \in \mathbb{R}^h \) is the final representation of \( L_i \). Then we can obtain the output of IKE as \( E_{<k} = [e_0, e_1, \ldots, e_{k-1}] \).

The IKE is aimed to acquire the embedding of each prediction step from the pre-built interdisciplinary graph \( G \). Fig. 4 shows an overview of IKE, which consists of multi-layer Graph Convolutional Network\(^4\) (GCN) and a Readout Layer. From an overall perspective, the label embedding is learned by:

\[ E_{<k} = \text{IKE}(L_{<k}, G), \]  

where the \( E_{<k} \in \mathbb{R}^{k \times h} \) is the matrix to preserve the prediction state and interdisciplinary knowledge from previous \( k - 1 \) steps.

Our Perspective: Embed the Previous Prediction. We usually model people’s behavior by capturing their relationships with others on the social network graph. The same idea is adopted in IKE to extract the previous prediction state from the interdisciplinary graph. Instead of inputting the former prediction result directly or passing a hidden state between each step, our design uses the previous results and their neighborhood on the interdisciplinary graph to obtain the representation of each step.

Step 1: Obtain the Representation. Suppose we are processing the \( k \)-th step. Given the sequence of label sets \( L_{<k} = [L_0, L_1, \ldots, L_{k-1}] \) from the previous \( k - 1 \) steps (or from a manually given partial label). Each label set \( L_i \) consists of the disciplines for holding the prediction result in the \( i \)-th step. We first treat the including discipline nodes as the central node and sample their \( N_g \)-hops neighborhoods (as an adjacency weighted matrix) from the interdisciplinary graph \( G \), where the \( N_g \) is the total layer number of GCN. Then we feed the adjacency weighted matrix and its corresponding node features into the GCN:\(^5\)

\[ H_i = N_g \times GCN(E_i, H^{(0)}_i), \]  

where the \( H^{(0)}_i \) is a random initialized input feature of the central nodes and their \( N_g \)-hops neighborhood nodes. The \( E_i \) is the weighted adjacency matrix between these nodes. The \( H_i \) is the output matrix undergoing \( N_g \) times GCN propagations.

Step 2: Readout the Features. Due to the varied number of discipline nodes within each prediction label set, we have to pool the matrix and acquire the representations of each step. For example, the representation of \( L_i \) is obtained by:

\[ e_i = \text{Readout}(H_i, L_i), \]

where the Readout layer is a lookup operation (i.e., take the \( L_i \) included node embedding) followed by a mean pooling layer. The \( e_i \in \mathbb{R}^h \) is the final representation of \( L_i \). Then we can obtain the output of IKE as \( E_{<k} = [e_0, e_1, \ldots, e_{k-1}] \).

\[ S_k = N_g \times IFBlock(E_{<k}, D), \]  

\(^4\)The detailed explanation of the Graph Convolutional Network can be found in [19].

\(^5\)We implemented the GCN with the message-passing design [20].
where $N_d$ is the total layer number of IF Block, and $S_k \in \mathbb{R}^{k \times h}$ is the fusion matrix that integrated the semantic information and the previous prediction information with interdisciplinary knowledge.

Our Perspective: Adaptively Utilize Each Part of Information.

There are two aspects we consider in IF. First, the current prediction state should be significantly affected by its previous results. Second, the model should adaptively choose the critical part of semantic information according to its current prediction state. To advance those, we divide the IF into two steps.

Step 1: Positional Encoding and Previous Label Embedding Fusion. We first sum the $E_{<k}$ with positional encoding to obtain $S_{<k}^{(0)}$ for preserving the order of prediction. Then, in layer-$l$ we perform Multi-Head Self-Attention between each prediction result representation to help each element in $S_{<k}^{(l-1)}$ aggregate knowledge from their context:

$$S_{<k}^{(l)} = S_{<k}^{(l-1)} \circ \text{MultiHead}(S_{<k}^{(l-1)}, S_{<k}^{(l-1)}, S_{<k}^{(l-1)}),$$

where $\circ$ operation is a Layer Normalization with a Residual Connection Layer. In this step, the MultiHead($\cdot$) adaptively aggregate the context information from every step prediction, and the $\circ$ operation integrate the context knowledge to the prediction result embedding and form the the level-$l$ prediction state as $S_{<k}^{(l)} \in \mathbb{R}^{k \times h}$. With the order information, we believe step 1 of IF can capture the hierarchical dependency and the interdisciplinary knowledge in $\mathbb{L}_{<k}$.

Step 2: Obtain the Semantic Information Adapatively. As mentioned before, we hope our model can utilize each part of the research proposal adaptively through prediction progress. Thus, we treat the prediction state $S_{<k}^{(l)}$ as Query and the representation of document set $D$ as Key and Value into another Multi-head Attention to propagate the semantic information:

$$Z^{(l)} = S_{<k}^{(l)} \circ \text{MultiHead}(S_{<k}^{(l)}, D, D),$$

$$S_{<k}^{(l)} = Z^{(l)} \circ \text{FC}(Z^{(l)}),$$

The MultiHead($\cdot$) adaptively aggregate the semantic information from the research proposal by the current prediction state, and the $\circ$ operation integrate the semantic information to the label embedding and construct the hidden feature $Z^{(l)} \in \mathbb{R}^{k \times h}$ in level-$l$. After a Fully-connected Layer denoted as $\text{FC}(\cdot)$ and the $\circ$ operation, we form $S_{<k}^{(l)}$ to hold the fusion information in level-$l$.

After $N_d$ times propagations, we acquire the $S_k = S_{<k}^{(N_d)}$ to represent the output of IF. With the IF, the model will learn the dependency between previous prediction steps and adaptively fuse each part of semantic information by the current prediction progress.

E. Level-Wise Prediction

The right side of Fig. 5 demonstrate the prediction in step-$k$. In LP, We feed the fusion feature matrix $S_k$ into a Pooling Layer, a Fully-connected Layer, and a Sigmoid Layer to generate each label’s probability for $k$-th level-wise label prediction.

$$y_k = \text{Sigmoid}(\text{FC}_k(\text{Pooling}(S_k))),$$

where the $y_k$ is the probability (i.e., $P(L_k|D, G, \gamma, \mathbb{L}_{<k}; \Theta)$ in (5) for each discipline in $k$-th level of hierarchical discipline structure $\gamma$. It is worth noting that we add a $l_{stop}$ token in each step to determine the prediction state. Thus, the $FC_k(\cdot)$ denotes a level-specific feed-forward network with a ReLU activation function to project the input to a $|C_k| + 1$ length vector, where represents the label number of level-$k$ with $l_{stop}$ token. After the $\text{Sigmoid}(\cdot)$, the final output $y_k$ is the probability of $k$-th level’s labels. Thus, the level-$k$’s objective function $\mathcal{L}_k$ can be defined as:

$$\mathcal{L}_k(\Theta) = \sum_{i=1}^{l_{stop}} (Y_k(i) \log(y_k(i)) + (1 - Y_k(i)) \log(1 - y_k(i))),$$

where $Y_k(i) = \left\{ \begin{array}{ll} 0 : & l_i \notin L_k \\ 1 : & l_i \in L_k \end{array} \right.$,

IV. Experiments

In this section, we conducted experiments to evaluate the performance of HIRPCN and answer the following questions: Q1. Is the performance of HIRPCN superior to the existing baseline models? Q2. How is each component of HIRPCN affect the
performance? Q3. Can HIRPCN achieve the best performance on the prediction at all levels? Q4. How good is the prediction result? Q5. Can the given partial expert advice improve the performance of HIRPCN? Q6. Can HIRPCN adaptively utilize the semantic information and historical prediction results? Q7. How is the word-level attention mechanism works for a better semantic extraction? Q8. What is the extra generated label significance to the peer-review assignment? We also evaluated the model performance from different settings of interdisciplinary graphs and hyperparameters.

A. Experimental Setup

1) Dataset Description: We collected research proposals written by the scientists from 2020’s NSFC research funding application platform, containing 280683 records with 2494 ApplyID codes. As shown in Table I, when the level goes deeper, since there is more number of disciplines, the classification becomes harder. In data processing, we chose four parts of a research proposal as the document data: 1) Title, 2) Keywords, 3) Abstract, 4) Research Field. The Abstract part is in long-text form, and the average length is 100. The rest three documents are all short text. We removed all the punctuation and padded the length of each document to 200. Then we added a type-token at the beginning of each document to mark its particular type (just like the [CLS] token in the Bert [21]). For the label construction, we generate the topic path of each research proposal (as described in Section II-A3). We further organized and divided those processed data by their interdisciplinarity into three datasets:

- **RP-all**: An overall dataset with all kinds of research proposals, whether interdisciplinary research or non-interdisciplinary research.
- **RP-bi**: Including all research proposals with two major disciplines or one major discipline but gets interdisciplinarity in its sub-discipline.
- **RP-differ**: Only includes the interdisciplinary research proposals with two major disciplines and its data shows the most interdisciplinarity among the three datasets.

In Table II, we counted the lengths of each label sequence and the average number of level-wise labels on three datasets. Generally speaking, from RP-all to RP-differ, the data exhibit interdisciplinarity increasingly and becomes more challenging to detect its topic path. The comparison of model performance between these datasets can evaluate the specific design of HIRPCN.

2) Baseline Algorithms: We compared our model with seven Text Classification (TC) methods including TextCNN [22], DPCNN [23], FastText [24], [25], TextRNN [26], TextRNN-Attn [27], TextRCNN [28], Trasformer [12] and three state-of-the-art HMC approaches including HMCN-F, HMCN-R [29], and HARNN [30]. We chose three Hierarchical Multi-label Text Classification (HMTC) methods including HE-AGCRCNN [31], HiLAP [32], two variants of HiAGM [33] (i.e., HiAGM-LA and HiAGM-TP). We also posted three ablation models of HIRPCN as baselines, including w/o Interdisciplinary Graph (w/o IGE), w/o Hierarchical Transformers (w/o HARNN), and w/o ALL. We summarized four critical attributes of these baseline models in Table III as:

- **Hierarchical Awareness**: represents whether the model uses the hierarchical information of the label structure to enhance the model performance.
- **Text-type Awareness**: represents whether the model is aware of the document type within each part of the research proposal.
- **Interdisciplinary Awareness**: represents whether the model can utilize the interdisciplinary knowledge in the discipline system, i.e., the interdisciplinary graph.
- **Start Prediction From Anywhere**: represents whether the model can start prediction with a given partial label, which is critical when aiding applicants and administrators in judging the proposal belongings.

3) Evaluation Metrics: To fairly evaluate the HIRPCN with other baselines, we adopted several metrics from the Multi-label Classification [34], [35], [36], i.e., the Micro-F1 (MiF1), and Macro-F1 (MaF1). We conducted a 5-fold cross-validation and reported the average recommendation performance.

4) Hyperparameters, Source Code and Reproducibility: We set the SIE layer number $N_e$ to 8, the dimension size $h$ to 64, and the multi-head number to 8. The IGE layer number $N_g$ is set to 1. The IF layer number $N_p$ is set to 8, and the multi-head number is set to 8. We pre-trained the word embedding with a dimension ($h$) 64 via Word2Vec [16] model. For the detail of training, we use Adam optimizer [37] with a learning rate of $1 \times 10^{-3}$, and set the mini-batch as 512, adam weight decay as $1 \times 10^{-7}$. The dropout rate is set to 0.2 to prevent overfitting. The

---

**TABLE I**

THE DISCIPLINE AND SUB-DISCIPLINE NUMBERS ON EACH LEVEL OF HIERARCHICAL DISCIPLINE STRUCTURE

| Prefix | Major Discipline Name | Total | $|C_2|$ | $|C_3|$ | $|C_4|$ |
|--------|------------------------|-------|-------|-------|-------|
| A      | Mathematical Sciences  | 318   | 6     | 57    | 255   |
| B      | Chemical Sciences      | 392   | 8     | 59    | 325   |
| C      | Life Sciences          | 801   | 21    | 162   | 618   |
| D      | Earth Sciences         | 166   | 7     | 94    | 65    |
| E      | Engineering and Materials Sciences | 138 | 13 | 118 | 7 |
| F      | Information Sciences   | 100   | 7     | 88    | 5     |
| G      | Management Sciences    | 107   | 4     | 57    | 46    |
| H      | Medicine Sciences      | 456   | 29    | 427   | 0     |
| -      | Total Disciplines      | 2478  | 95    | 1042  | 3321  |

**TABLE II**

THE DETAILS OF THREE DATASETS

| Dataset          | RP-all | RP-bi | RP-differ |
|------------------|--------|-------|-----------|
| Total Proposal Number | 280683 | 125466 | 20632     |
| #Avg. Labels Length | 2.393  | 3.355 | 3.397     |
| #Avg. Labels Num in Level-1 | 1.073  | 1.164 | 2.000     |
| #Avg. Labels Num in Level-2 | 1.197  | 1.442 | 2.000     |
| #Avg. Labels Num in Level-3 | 1.364  | 1.870 | 1.985     |
| #Avg. Labels Num in Level-4 | 0.477  | 0.726 | 0.808     |

---

4https://github.com/RingBDStack/HE-AGCRCNN
3https://github.com/morningmoni/HiLAP
5https://github.com/RingBDStack/HE-AGCRCNN

---

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
warm-up step is set as 1000. The $\alpha$ and $\beta$ for interdisciplinary graph construction set to 0.1. We shared the source code via Dropbox.9

5) Environmental Settings: All methods are implemented by PyTorch 1.8.1 [38]. The experiments are conducted on a CentOS 7.1 server with an AMD EPYC 7742 CPU and 8 NVIDIA A100 GPUs.

B. Experimental Results

1) RQ1: Overall Comparison: We evaluated every method on all datasets and reported the results in Table III. To compare the overall performance of the different models, we organized each step’s prediction results flatly. Then, we used the MiF1 and the MaF1 as the overall evaluation metrics. From the results, we can observe that: First, the HIRPCN achieved the best performance among all datasets, proving that HIRPCN could better solve the challenges mentioned in the IRPC task. Besides this, the methods with top-down HMC schema (e.g., HMCN, HARNN, HIRPCN, and its variants) universally performed better than TC methods. The first reason is that the methods with top-down HMC schema organize the labels in a hierarchical view rather than a flat view, which can reduce the candidate label number in each step. Second, one of the critical attributes of the top-down HMC schema is that the model will generate the current prediction based on its previous prediction state. Because there are fewer categories in the earlier levels, the classification in these levels will be easier to predict correctly. Thus, the precise former results will facilitate the performance of these methods in the rest steps. We can also observe that the HMTC methods (e.g., HE-AGCRNN, HiLAP, and HiAGM) performed better than the TC method but slightly worse than the HMC methods. The reason for the former is that their encoder is similar to the TC methods, the HMTC methods either adopted the hierarchical structure into loss function (e.g., the HE-AGCRNN), initialized the label embedding by hierarchical structure (e.g., the HiAGM and its variants), or explicitly using the dependency of each label (e.g., the HiLAP). The latter is because we adopted the Hierarchical Transformer-based encoder for semantic information extraction in these HMC methods, which is more suitable for modeling long text. Further, we noticed that the model’s performance on the three datasets gradually declined. As we mentioned before, the interdisciplinary proposals have more labels on each level, making the classification task harder. The performance on RP-bi is better than that on RP-differ, indicating due to the disparity between the two major disciplines, identifying the topic path under them will become much more challenging. Also, the performance of HIRPCN is better than other baselines, showing its robustness and generalization on the datasets with different interdisciplinarity.

2) RQ2: Ablation Study: We also discussed the performance of each ablation variant of HIRPCN. The w/o SIE replaced the SIE component with a vanilla Transformer, making the model unable to discriminate the different types of documents (no Text-type Awareness). The w/o IKE removed the IKE component. In each prediction step, w/o IKE will initialize the input previous prediction results by a look-up table (no Interdisciplinary Awareness). The w/o ALL ablated both the IKE and SIE to represent the blank control group. From the Fig. 3, we can observe that: First of all, the HIRPCN beats all its ablation variants. This observation proves that both SIE and IKE components are valid in our model design. We also noticed that w/o SIE performs worse than w/o IKE in each dataset (e.g., from $-3.8\%$ to $-8.2\%$ deterioration on Micro-F1). This phenomenon indicates that the type-specific semantic information will play a more critical role in topic path detection. Second, w/o IKE performed slightly worse (e.g., $-1.5\%$ deterioration on Micro-F1) than HIRPCN on the RP-all dataset, while this deterioration (e.g., $-1.9\%$ and

| Dataset | Hierarchical Awareness | Text-type Awareness | Interdisciplinary Awareness | Start Prediction From Anywhere | RP-all | RP-bi | RP-differ |
|---------|------------------------|---------------------|---------------------------|--------------------------------|--------|--------|-----------|
| TextCNN | x                      | x                   | x                         | x                              | 0.456  | 0.168  | 0.426     | 0.148 | 0.438  | 0.063     |
| DPCNN   | x                      | x                   | x                         | x                              | 0.376  | 0.108  | 0.364     | 0.080 | 0.346  | 0.020     |
| FastText| x                      | x                   | x                         | x                              | 0.460  | 0.167  | 0.427     | 0.149 | 0.432  | 0.067     |
| TextRNN | x                      | x                   | x                         | x                              | 0.402  | 0.100  | 0.364     | 0.077 | 0.351  | 0.021     |
| TextRNN-Attn | x | x | x | x | 0.413 | 0.098 | 0.367 | 0.072 | 0.338 | 0.018 |
| TextRCNN| x                      | x                   | x                         | x                              | 0.426  | 0.138  | 0.381     | 0.100 | 0.367  | 0.025     |
| Transformer | x | x | x | x | 0.370 | 0.083 | 0.339 | 0.066 | 0.338 | 0.023 |
| HMCN-F  | ✓                      | x                   | x                         | x                              | 0.676  | 0.459  | 0.582     | 0.308 | 0.569  | 0.161     |
| HMCN-R  | ✓                      | x                   | x                         | x                              | 0.599  | 0.289  | 0.506     | 0.185 | 0.478  | 0.069     |
| HARNN   | ✓                      | ✓                   | x                         | x                              | 0.686  | 0.443  | 0.574     | 0.294 | 0.516  | 0.102     |
| HE-AGCRNN| ✓                  | ✓                   | x                         | x                              | 0.513  | 0.331  | 0.467     | 0.289 | 0.463  | 0.118     |
| HI-LAP  | ✓                      | x                   | x                         | x                              | 0.628  | 0.418  | 0.536     | 0.280 | 0.528  | 0.141     |
| HIAGM-LA| ✓                      | ✓                   | x                         | x                              | 0.594  | 0.400  | 0.505     | 0.258 | 0.501  | 0.134     |
| HIAGM-TP| ✓                      | ✓                   | x                         | x                              | 0.629  | 0.428  | 0.548     | 0.285 | 0.535  | 0.135     |
| w/o ALL | ✓                      | x                   | x                         | x                              | 0.704  | 0.429  | 0.583     | 0.259 | 0.512  | 0.098     |
| w/o SIE | ✓                      | x                   | ✓                         | ✓                              | 0.710  | 0.438  | 0.587     | 0.276 | 0.533  | 0.099     |
| w/o IKE | ✓                      | ✓                   | ✓                         | ✓                              | 0.732  | 0.448  | 0.636     | 0.328 | 0.577  | 0.127     |
| HIRPCN  | ✓                      | ✓                   | ✓                         | ✓                              | 0.748  | 0.512  | 0.648     | 0.384 | 0.635  | 0.180     |

9https://www.dropbox.com/sh/x5m1jlcax8jp6tk/AAAg-KrGM8kHZuqoSFHKLRc0a

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
and hierarchy between each level.

Table IV showed the level-wise Micro-F1 and level-wise Macro-F1 classification results on four levels. From the results, we can observe that HIRPCN outperforms all the baseline methods and the variants on every level, showing the advanced ability of classifying at all granularity. We also observed that the performance of each model tends to decrease with the depth of level increase. The reason is that the number of categories increases rapidly, making the classification task harder. We also noticed that Macro-F1 dropped relatively more suddenly than Micro-F1 on each method. From the dataset statistic, we interpreted this as: 1) the length of the label path is variable, so in some cases, the last two levels might contain no target. 2) the test set is imbalanced, so some disciplines might not contain samples. Thus, the Macro-F1 will be lower than Micro-F1, especially in the last two levels. The level-wise model performance on RP-all and RP-bi show the same tendency as on RP-differ so we won’t go into detail.

### 3) RQ3: Level-Wise Performance:

We reported the performance of level-wise prediction of HIRPCN and other baselines on the RP-differ dataset. Table IV showed the level-wise Micro-F1 and level-wise Macro-F1 classification results on four levels. From the results, we can observe that HIRPCN outperforms all the baseline methods and the variants on every level, showing the advanced ability of classifying at all granularity. We also observed that the performance of each model tends to decrease with the depth of level increase. The reason is that the number of categories increases rapidly, making the classification task harder. We also noticed that Macro-F1 dropped relatively more suddenly than Micro-F1 on each method. From the dataset statistic, we interpreted this as: 1) the length of the label path is variable, so in some cases, the last two levels might contain no target. 2) the test set is imbalanced, so some disciplines might not contain samples. Thus, the Macro-F1 will be lower than Micro-F1, especially in the last two levels. The level-wise model performance on RP-all and RP-bi show the same tendency as on RP-differ so we won’t go into detail.

### 4) RQ4: Evaluation of Prediction Results:

The Micro-F1 and the Macro-F1 can measure how accurate the prediction is. However, it might be insufficient to evaluate the HIRPCN in real-world applications. In this experiment, we conducted two statistical analyses on three datasets for prediction results. The first analysis evaluates how close the prediction is to the truth label. For instance, if given a truth label in the 3rd level as \(\{A0101\}\), a prediction \(\{A0102\}\) might be more acceptable than \(\{B0101\}\), although they are both wrong. Based on that, we developed a metric named interdisciplinary distance given as:

\[
D(L_k, \hat{L}_k) = \sum_{i \in L_k} \min \left( \left\{ \text{dis}(l, \hat{l}) \right\}_{l \in \hat{L}_k} \right),
\]

where \(L_k\) and \(\hat{L}_k\) are the truth labels and the prediction results in level-\(k\), respectively. \(\min(\cdot)\) will take the minimized value from a given set. We adopted a revised edit distance as the distance function \(\text{dis}(\cdot, \cdot)\). In detail, if any two-digit or major discipline in the predicted ApplyID code is different from the truth, the function will instantly return the penalty. By discussing it with funding agencies, we set the penalty as \(\{1, 10, 30, 50\}\), depending on the remaining step. For example, the distance between \(A0101\) and \(A0102\) will be 1 due to they are different in level-3, and the remaining step will be 0. And the distance between \(A0101\) and \(B0101\) will be 30 because they are already different in the first level and their remaining step will be 2. Also, if \(l_1\) or \(l_2\) is null (i.e., the \(L_k\) or \(\hat{L}_k\) is empty), the distance will be 0. Thus, \(D(L_k, \hat{L}_k)\) will quantitatively assess the distance between truth label set \(L_k\) and predicted set \(\hat{L}_k\). Then, we level-wise average the distance of every sample from the test set. The higher average interdisciplinary distance represents the prediction for this level-\(k\) are more offset. Another analysis is we count the appearance number of three wrong cases, i.e., \(\text{Lack}, \text{TooMuch}, \text{Wrong}\), and \(\text{Other}\). \text{Lack} represents the prediction in the current level is null, which means the model stop at a coarse level. \text{TooMuch} means the model stop into a too-fine-grained level, and the truth label in the current level is null. \text{Wrong} represents that the model gives a prediction that not following the hierarchy. \text{Other} represents all other wrong cases.

We calculated the average distance of each level and counted the number of wrong cases. The results of these analyses were reported in Fig. 6. First, we can observe that the interdisciplinary distance and the number of wrong cases grow from level-1 to level-4, which is the same as the observation of the decline of Micro-F1 and Macro-F1 in Table III. Second, HIRPCN outperforms the second and third-best methods in the distance and shows excellent superiority in the first few levels. This observation shows the same phenomenon as the comparison with Micro-F1 and Macro-F1 in Table III. Furthermore, due to the most happening wrong cases in the level-4 of RP-differ is \text{Lack}, this distance is caused by the model stopping prediction prematurely rather than giving an incorrect prediction, which is acceptable in the real world. Apart from those, we noticed that the most reason to cause the distance between predictions and truth labels is \text{Lack}. Simultaneously, the appearance number of \text{TooMuch} is relatively less, showing that our model will be more prude when predicting the topic path. Additionally, \text{Wrong} happened around 0.7% to 4.0% in percentage in each level, showing that although we did not provide an explicit coherence constraint, the HIRPCN will instinctively learn the dependency and hierarchy between each level.

### 5) RQ5: Prediction With Given Labels:

The Given Labels is a partial label set provided by applicants or funding administrators when they try to fill the ApplyID. As mentioned in the previous section, we hope the HIRPCN can be an assistant who knows the
9694 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 35, NO. 9, SEPTEMBER 2023

Fig. 6. The details of the wrong case and level-wise prediction distance on three datasets. The attached table also reported the interdisciplinary distances of the second and third-best methods.

Fig. 7. The performance and improvement with given labels in terms of Micro-F1.

whole picture of the discipline system to aid the categorization with the manager. This study evaluated the improvement when HIRPCN receives incomplete labels and predicts the rest label. It is also worth noting that other baseline methods use a hidden vector to represent previous prediction results, which cannot initialize the prediction from a given partial label set. Thus, we compared the performance of HIRPCN with different levels of provided expert knowledge on the RP-differ dataset. As you can see in Fig. 7, the first row of this heatmap showed the Micro-F1 of the HIRPCN on RP-differ when applicants were not given the partial labels. The rest rows illustrated when the first level, first and second level, first, second and third level were given and how the model improved its performance using these partial labels. The color of each cell showed the strength of the improvement. From the heatmap, we can observe that: First, with the incomplete labels given, the overall performance of HIRPCN rises. This phenomenon proves that the architecture enables the model to utilize the partial label to improve the overall performance. Second, we noticed that the given partial labels could significantly improve the next-level prediction but decline in the remaining level. We interpret this phenomenon as the current prediction result might be more related to its most recent-level results. So, the model would be trained to raise a high attention on the most current level rather than other earlier-level. In summary, the improvement from the given partial label shows the potential application scenario for the AI assistant system.

6) RQ6: Attention Mechanism on Information Fusion: Adaptively utilizing each part of the research proposal and the previous prediction results is the most critical idea of the Information Fusion component. In this experiment, we gathered the model’s attention value on each previous prediction step and each different type of text from the IF component. The results was reported in Figs. 8 and 9.

The Attention to Different Types of Text. Fig. 8 illustrated the attention value of the different types of documents in each step. We observed that the Title field is important while predicting the major discipline (step 1). However, HIRPCN will pay more attention to the Research Field and Abstract to generate the rest labels. Although, it might not be noteworthy to interpret and compare the model mechanism with experts’ behavior. But, compared to simply concatenating the text sequence, the attention mechanism could provide more potential for information fusion.

The Attention to Historical Prediction Results. Fig. 9 showed the model’s attention to the historical prediction results. We noticed that the model would adopt the interdisciplinary knowledge information from the most recent predictions. However, the model might consider other historical predictions to generate the rest in some cases. This phenomenon also proves the experiment observation in RQ5 (i.e., the given partial label will significantly improve the next level’s performance).

To sum up, the attention in the Information Fusion will adaptively variate during the prediction progress. This phenomenon shows the validity of the model component and brings excellent potential to the HIRPCN.

7) RQ7: Attention Mechanism on Text: Fig. 10 is a case study of an interdisciplinary research proposal. If a word token gained more attention from SIE, it would be shaded redder in
the self-attention mechanism worked well in the documents with multiple types and variable lengths, such as the term “耐药相关” (Drug Resistance) in the Title or the words “挖掘” (Mining), “算法” (Algorithm) in the Abstract. We believe this phenomenon shows why the SIE is superior for handling the multiple texts in research proposals. In addition, those highlighted critical words can bring interpretability to HIRPCN, which is also essential for fund administrators.

8) RQ8: Hidden Interdiscipline Find: Among all the wrong cases in RP-all, the HIRPCN classifies part of non-interdisciplinary research to interdisciplinary, which means except for their labeled discipline, our model predicts an extra discipline. After observing those wrong cases, we noticed that these non-interdisciplinary research proposals were somewhat related to their extra predicted disciplines. To understand this phenomenon, We invited eight fund administrators from NSFC with an insightful understanding and management experience of the eight major disciplines to judge if those extra disciplines can provide helpful information for improving the reviewer assignment. We first group those cases by the extra discipline predicted by HIRPCN. Then we sample 50 research proposals from each group. We send those research proposals to each expert according to the expert’s familiar discipline. If the expert thinks the reviewers from the extra discipline should be considered to supplement the peer-review group, they will mark it as correct, otherwise wrong. We illustrated the percentage of correct of each extra discipline in Fig. 11. From the results, we can observe that those experts marked 57% of the samples as correct. Further, all experts commented that the HIRPCN could detect the hidden interdisciplinary research domain. Besides these, some experts commented that although some cases were marked wrong, most of them explicitly use the keywords from the predicted extra discipline. The Experts categorized those cases into non-interdisciplinary research because their focal
Fig. 11. The results of hidden interdiscipline find study.

Fig. 12. Four cases in different area, while HIRPCN detect another hidden discipline.

point is on the labeled discipline instead of the predicted extra one (e.g., a study on Artificial Intelligence might introduce the ideas and terms from Neural Sciences. Still, the neuroscientists might be not suitable to review it). Furthermore, some experts commented that they marked some cases as wrong because the investigators created the labeled discipline to cover the extra discipline topic. So, the labeled discipline is enough to describe the research domain. For example, the discipline Bionics and Artificial Intelligent (C1005) in Life Sciences (C), was created to represented the interdiscipline of Bionics (C10) and Artificial Intelligent (F06).

We randomly picked four proposals and listed them in Fig. 12. Applicants manually labeled those proposals with one discipline, but our model extra marked them with another domain from the Information Sciences. We can observe that those research proposals are highly related to their labeled disciplines, while they also strongly bond with the Information Sciences, whether the methodology or the idea. We believe that, with the HIRPCN, those research proposals can be reviewed by experts who are knowledgeable in both areas, thus providing a fair evaluation.

C. Interdisciplinary Graph Settings

As mentioned in Section II-A4, we introduced a weighted graph to represent the interdisciplinarity between each discipline. This weight is defined by the proportion (disparity) and the citation number (frequency) of the overlap topic in (3). On the one hand, if two domains show high interdisciplinarity, their research area will have great distinction overall, but their overlap topic will be a significant hot area to research. On the other hand, two disciplines that have too many overlapped topics or their overlap topic have little citation might be two similar disciplines (e.g., the Machine Learning and the Deep Learning) or two irrelative areas (e.g., the Management Sciences and the Electronic Engineering), both show little interdisciplinarity. We use two parameter $\alpha$ and $\beta$ to set the weight of two component.

We trained the HIRPCN on RP-differ with different combinations of $\alpha$ and $\beta$. The result of model performance in epoch 100 is visualized in Fig. 13. From the figure, we can observe that the best setting of the interdisciplinary graph is set the $\alpha$ and $\beta$ to 0.1. If we both set the $\alpha$ and $\beta$ to 0, all edge weights will equal 1.0, and the graph will deteriorate to a co-topic graph. In epoch 100, this setting would undermine the model performance by around $-3.5\%$ compared to the best setting. We believe this phenomenon will explain the importance of interdisciplinary knowledge and our adopted RS measurement.

D. Hyperparameter Selection

We conducted experiments to evaluate the effect of four key hyperparameters, i.e., $N_e$, $N_d$, $N_g$, and the number of self-attention heads in the SIE and IF. We investigated the sensitivity of those parameters in each part of HIRPCN and reported the results in Fig. 14. From the figure, we can observe that with the $N_e$ and $N_g$ vary from 1 to 8, the performance of HIRPCN increases at first and then decreases slowly. The model will likely be over-fitting to impact the classification results with more parameters introduced. Also, we found that the model would have the best performance when IKE sampled a one-hop neighborhood to extract interdisciplinary knowledge. The reason is that the interdisciplinary graph has many edges, and two or more hop sampling strategies will make the sampled sub-graph denser, which will provide unrelated knowledge for IKE. Lastly, more attention heads will generally improve the
performance of HIRPCN. Meanwhile, we found that too many attention heads will make the model perform worse. Based on the experimental result, we set $N_e$ to 8, $N_d$ to 8, $N_p$ to 1, and the number of self-attention heads to 8 to achieve the best performance.

V. RELATED WORKS

A. Text Classification

The text classification [39] problem is a classical natural language processing task, which aims to extract intrinsic semantic information and assign its related labels. The shallow approaches, such as Word2Vec [16], and GloVe [40], aim to map the word sequence into vector space and then classify each document by this representation. The deep-learning approaches developed many encoders to enhance the semantic information extraction, such as RNN-based [26], [27], [41], CNN-based [22], [23], [28], Attention-based [12], and pretrained language model based [21]. Apart from these approaches, many solutions have been proposed to solve the Multi-label Text Classification, such as WRAP [42], LCBM [43], and 3WDLE [44]. However, in a real-world scenario, the labels of text classification are always exhibit a hierarchical structure, which named the Hierarchical Multi-label Text Classification (HMTC) task. To address this challenge, the study [31] learns the label dependency and uses this dependency to construct the loss function. HiLAP [32] proposed a reinforcement approach so the prediction can follow the label structure. HiAGM [33] proposed two structure encoders for modeling the hierarchical labels. HiMatch [45] implemented a text-to-hierarchical-label matching network. The interdisciplinary research proposal classification task can be seen as a complicated form of HMTC task which aims to categorize multiple types of documents into a hierarchical discipline structure.

B. Hierarchical Multi-Label Classification

Our work is most related to deep-learning based hierarchical multi-label classification [46] (HMC) methods which leverage standard neural network approaches for multi-label classification problems [29], [47], [48] and then exploit the hierarchy constraint [30] in order to produce coherent predictions and improve performance. Zhang et al. [49] propose a document categorization method with hierarchical structure under weak supervision. The work [50] design a attention-based graph convolution network to category the patent to IPC codes. The work in [51] use the convolution neural network as an encoder and explore the HMC problem in image classification. The works in [52] propose a active learning approach for HMC problem. Aly et al. [53] propose a capsule network based method for HMC problem. TAXOGAN [54] embedding the network nodes and hierarchical labels together, which focuses on taxonomy modeling. In recent studies [55], researchers try to exploit more information among the label embedding and label correlation. Besides the label embedding approach, the study [56] explore the performance of hierarchy dependency as a training constraint and achieve well performance on a small dataset with neural networks. The previous studies in this area are well constructed and have significant success in a specific dataset. However, few studies focus on the complex sci-tech text data classification with a hierarchical discipline structure. In this article, we inherit the fundamental idea of HMC networks to construct our novel classification methods on the hierarchical discipline structure.

VI. CONCLUSION

This article proposed HIRPCN, a novel HMC method for the IRPC task on the real-world research proposal dataset. This method extracts the semantic information from each document separately and combines them to achieve a high-level abstraction. Then, the model process each previous level’s prediction result and neighborhood topology structure on the pre-defined interdisciplinary graph as interdisciplinary knowledge. The prediction step integrates each part of the research proposal by attention mechanism and generates the next level prediction. The experiments show that our model achieves the best performance on three real-world datasets and can provide the best granular on each level prediction. Other experiments explains the attention mechanism and points out that our model could fix the incomplete interdisciplinary labels under a domain-specific expert evaluation. With the ability to start prediction from any given label, HIRPCN could assist whether researcher or fund administrator to fill the discipline information, which is crucial to improve the reviewer assignation.

REFERENCES

[1] S. Milojević, “Quantifying the cognitive extent of science,” J. Informetrics, vol. 9, no. 4, pp. 962–973, 2015.
[2] S. Fortunato et al., “Science of science,” Science, vol. 359, no. 6379, 2018, Art. no. eaao0185.
[3] J. G. Foster, A. Rzhetsky, and J. A. Evans, “Tradition and innovation in scientists’ research strategies,” Amer. Sociol. Rev., vol. 80, no. 5, pp. 875–908, 2015.
Meng Xiao received the BS degree from the East China University of Technology, and the master’s degree from the China University of Geosciences, Wuhan. He is currently working toward the PhD degree with the University of Chinese Academy of Sciences. His main research interests include data mining, graph representation learning, hierarchical multi-label text classification combined with the knowledge, and the knowledge graphs of the discipline system.

Ziyue Qiao received the BE degree from the Wuhan University, China, in 2017. He is currently working toward the PhD degree with the Computer Network Information Center, University of Chinese Academy of Sciences. His research interests include data mining, graph learning, and natural language processing, with an emphasis on designing new algorithms for graph representation/transfer learning and academic data mining.

Yanjie Fu received the BE degree from the University of Science and Technology of China, in 2008, the ME degree from the Chinese Academy of Sciences, in 2011, and the PhD degree from Rutgers, the State University of New Jersey, in 2016. He is an assistant professor with the Department of Computer Science, University of Central Florida. His research interests include data mining and big data analytics.

Hao Dong received the BS degree from Shaanxi Normal University, China, in 2019. He is currently working toward the PhD degree with the Computer Network Information Center, University of Chinese Academy of Sciences. His research probes data mining, natural language processing, knowledge-based graph, and temporal graph.

Yi Du received the PhD degree from the Institute of Software, Chinese Academy of Sciences, in 2013. He is a professor with the Department of Big Data Technology and Application Development, Computer Network Information Center, Chinese Academy of Sciences. His research interests include big data and visual analytics.

Pengyang Wang received the PhD degree in computer science from the University of Central Florida. He is an assistant professor with the State Key Laboratory of Internet of Things and Smart Cities, University of Macau. He has broad interests in data mining and machine learning, especially in geospatial-temporal data modeling. He has published in related domains on top venues such as KDD, IEEE Transactions on Knowledge and Data Engineering, IJCAI, WWW, AAAI etc. He also served as a (senior) program committee member among top conferences in the data mining and artificial intelligence communities.

Hui Xiong (Fellow, IEEE) received the PhD degree in computer science from the University of Minnesota, USA. He is currently a chair professor with the Hong Kong University of Science and Technology (Guangzhou). His research interests include data mining, mobile computing, and their applications in business. For his outstanding contributions to data mining and mobile computing, he was elected an AAAS fellow and an IEEE fellow in 2020.

Yuanchun Zhou received the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2006. He is a professor, PhD supervisor, the deputy director of Computer Network Information Center, Chinese Academy of Sciences. His research interests include data mining, big data processing, and knowledge graph.