Pre-Trained Language Models with Topic Attention for Supervised Document Structure Learning

Dang Pham
Department of Computer Science
New Mexico State University
dangpnh@nmsu.edu

Tuan M. V. Le
Department of Computer Science
New Mexico State University
tuanle@nmsu.edu

***Abstract*—The discourse-level structure of a document can be captured through learning the rhetorical functions of sentences in that document. Existing supervised methods based on pre-trained language models for classifying rhetorical functions of sentences usually focus on utilizing rhetorical words but ignore the topics of sentences. Since topic words can provide additional information for enhancing the learning of the document structure, we present a neural topic model that is integrated with a BERT-based language model through a unified probabilistic generative process for learning both the rhetorical structure and topic structure of documents. For inference, we design a topic attention mechanism to utilize the learned topic words from previous sentences to improve the prediction of the current sentence's rhetorical label. The extensive experiments on four real-world datasets of different domains show that the proposed model improves the detection of rhetorical functions of sentences and is effective in document modeling and extracting coherent topics.**

***Index Terms*—document structure learning, pre-trained language models, topic models, attention mechanism.**

I. Introduction

Document structure learning aims to detect rhetorical roles, functions, or intents of sentences in a document. For example, in a scientific abstract, each sentence can be assigned to one of rhetorical functions such as Background, Objective, Methods, Results, Conclusions. Organizing sentences into their rhetorical roles is important for document understanding [1], [2] and can help improve downstream tasks such as document segmentation [3], [4], argumentation mining [5], [6], and summarization [7]–[9].

The automatic identification of rhetorical roles or intents of sentences in a document can be modeled as a sequential sentence classification problem where each sentence is assigned an intent label taking into account the context from neighboring sentences [10], [11]. Existing methods typically propose different hierarchical sentence encoders that derive contextualized sentence representations based on each sentence's sequence of word embeddings and the sequence of sentences [10]–[13]. These methods have shown that utilizing contextualized representations of all words to jointly encode sentences in a sequence can improve the intent label prediction performance significantly.

Besides the sequence contextual information, topics of words in the sentence may carry additional information to improve the task. Words in a document can be divided into two types which are intent words and topic words [14]. For example, as shown in Figure 1, the first sentence in the abstract contains the intent words "purpose", "determine", "able", which clearly indicates the intent to describe "Objective" of this sentence. While intent words can be good indications of the sentence's intent, topic words can also provide additional information. Sentences that describe the same topic usually have the same intent and the topics of previous sentences may also contribute to the determination of the current sentence's intent. Therefore, analyzing the topic words of a sentence and its surrounding sentences may help improve further the prediction performance.

Existing supervised methods based on pre-trained language models for classifying rhetorical functions of sentences usually focus on utilizing rhetorical words but ignore topics of sentences. They do not explicitly recognize those two word types and utilize topics of words in the sentence. In contrast, although topic models such as LDA [15] can extract topics, they do not model intents at sentence level as well as the sequence nature of words and sentences in the document. Therefore, their performance for predicting sentence intents may be limited because they do not exploit contextualized representations of words and sentences. While jointly modeling intents and topics is possible as shown in [14], that work does not employ the contextual information of sentence representations learned through language models. Therefore, it does not inherit the strengths of language models in modeling sequence contextual information for predicting intents of sentences.

To inherit the strengths of both language models and topic models for document intent structure learning, we propose a joint supervised approach that integrates a sentence sequence model and a neural topic model for learning both the rhetorical structure and topic structure of documents. Our proposed model, called INTENT, includes a BERT-based language model that encodes all sentences in a sequence [16], and these sentence representations are then transformed to probability vectors through a feed-forward network for predicting sentence intents. We integrate the sentence sequence model with a neural topic model by coupling the generation of sentence intents with the generation of words. In our proposed generative process, the intent label of the current sentence will depend on the BERT representations and topics of previous sentences. After the current sentence's intent label is drawn,

979-8-3503-6248-0/24/$31.00 ©2024 IEEE

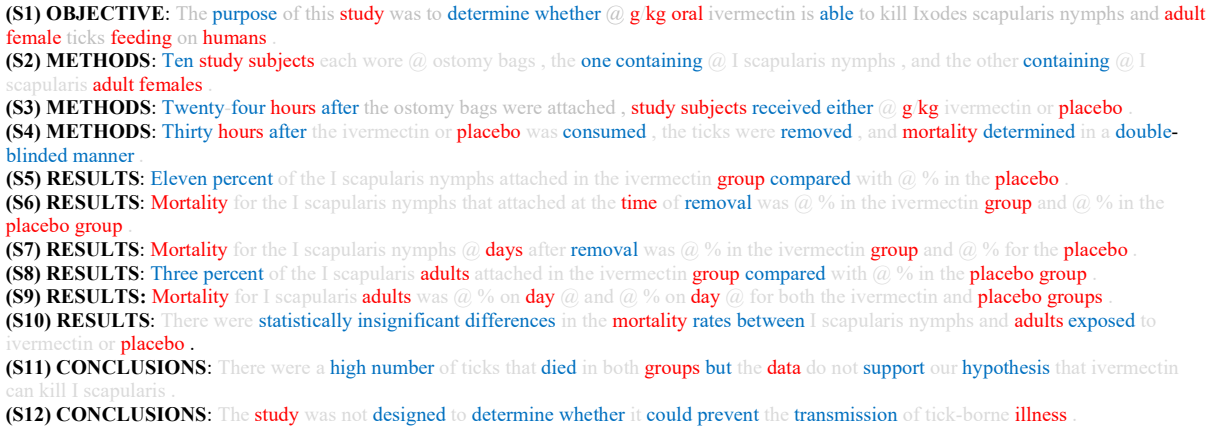


Fig. 1: Intent and topic words (in blue and red respectively) of a scientific abstract in PubMed20k detected by INTENT. The gray words are either stop words or words with low frequency. They are removed in the preprocessing step.

each word is generated depending on its word type (i.e., intent or topic words), intent, and topic. We use a random binary variable with a Bernoulli prior to model the type of words. Finally, to utilize topic information of previous sentences, we propose a topic attention model where it will learn how topic representations of previous sentences will contribute to the intent prediction of the current sentence. To the best of our knowledge, our proposed model is the first to jointly integrate a sentence sequence model and a neural topic model for supervised learning of document structure. We summarize our contributions as follows:

- We propose a neural probabilistic model, called INTENT[1], that integrates a sentence sequence model with a neural topic model for learning intent and topic structures of documents.
- While the sentence sequence model can capture dependencies and coherence in the output label sequence, we also propose a topic attention model that utilizes the topics of previous sentences for improving the intent prediction performance.
- We conduct extensive experiments on four real-world datasets of different domains. The results show that our proposed model is effective in detecting rhetorical functions of sentences, document modeling, and extracting coherent topics.

II. INTENT: A Supervised Model for Document Structure Learning

In this section, we present how the sentence sequence model, the neural topic model, and the topic attention model are integrated through a unified generative process. We also derive the variational inference algorithm of the proposed model INTENT.

A. Generative Model of INTENT

Given a corpus of D documents, $\mathcal{D} = \{w_1, \ldots, w_D\}$, where a document w_d is a sequence of N_d sentences, $w_d = (w_{d1}, w_{d2}, \ldots, w_{dN_d})$. A sentence w_{ds} is represented as a bag of words over the vocabulary of size V. K is the number of topics. $\mathcal{I}$ is the number of intents. The goal of document structure learning is to learn: 1) word distributions of intents, $\boldsymbol{\beta} = \{\beta_i \in \mathbb{R}^V\}_{i=1}^{\mathcal{I}}$; 2) word distributions of topics, $\boldsymbol{\psi} = \{\psi_k \in \mathbb{R}^V\}_{k=1}^{K}$; 3) topic distributions of documents, $\boldsymbol{\theta} = \{\theta_d \in \mathbb{R}^K\}_{d=1}^{D}$; 4) and the intent label of each sentence in the document. To infer all those parameters, we assume the following process to generate a document and the intents of all sentences in the document.

For each document d,

1) Draw topic distribution $\theta_d \sim \mathcal{LN}(0, I)$
2) For each sentence w_{ds} in document w_d:
   a) Obtain the representation h_{ds} of w_{ds} through a BERT-based language model and topic attention mechanism, conditioned on previous sentences $w_{d\{1:s-1\}}$
   b) Draw $l_{ds} \sim \text{Multi}(\text{softmax}(\text{MLP}(h_{ds})))$
   c) For each word w_{dsm}:
      i) Draw $b_{dsm} \sim \text{Bernoulli}(\gamma)$
      ii) if $b_{dsm} = 1$ (w_{dsm} is an intent word):
         A) Draw intent word $w_{dsm} \sim \text{Multi}(\beta_{l_{ds}})$
      iii) if $b_{dsm} = 0$ (w_{dsm} is a topic word):
         A) Draw topic $k_{dsm} \sim \text{Multi}(\theta_d)$
         B) Draw topic word $w_{dsm} \sim \text{Multi}(\text{softmax}(\mathcal{V}^\top T_{k_{dsm}}))$

In Step 1 of the generative process, $\mathcal{LN}(.)$ denotes the logistic-normal distribution. It transforms a Gaussian random variable to a variable on the topic simplex. The topic distribution θ_d is drawn from this distribution as follows:

$$x_d \sim \mathcal{N}(0, I);\ \theta_d = \text{softmax}(x_d) \tag{1}$$

Each topic k is represented as a vector $T_k \in \mathbb{R}^E$ and its word distribution ψ_k is computed as: $\psi_k = \text{softmax}(\mathcal{V}^\top T_k)$ where $\mathcal{V}$ is the word embedding matrix, $\mathcal{V} \in \mathbb{R}^{E \mathbf{x} V}$, and E is the embedding dimension. For each word w_{dsm} in sentence w_{ds}

[1] https://github.com/dangpnh2/INTENT

of document d, in Step 2(c)i., we draw its word type b_{dsm}:

$$b_{dsm} \sim \text{Bernoulli}(\gamma) \tag{2}$$

If it is an intent word, w_{dsm} is drawn from $\text{Multi}(\beta_{l_{ds}})$ where $\beta_{l_{ds}}$ is the word distribution of intent l_{ds}. If it is a topic word, a topic k_{dsm} is drawn and w_{dsm} is then generated from $\text{Multi}(\psi_{k_{dsm}})$, where $\psi_{k_{dsm}} = \text{softmax}(\mathcal{V}^\top T_{k_{dsm}})$ which is the word distribution of topic k_{dsm}:

$$k_{dsm} \sim \text{Multi}(\theta_d) \tag{3}$$

$$w_{dsm} \sim \text{Multi}(\text{softmax}(\mathcal{V}^\top T_{k_{dsm}})) \tag{4}$$

The parameters of intent and topic model including word distributions of intents ($\boldsymbol{\beta}$), word distributions of topics ($\boldsymbol{\psi}$), and topic distributions of documents ($\boldsymbol{\theta}$) can be learned by variational autoencoding. In Section II-D, we present a neural variational inference to infer those parameters.

B. Sentence Sequence Model for Generating Intents

In Steps 2(a) and 2(b) of the generative process, we introduce a BERT-based sentence sequence model to generate the intent label of each sentence in a document. For each sentence w_{ds} of document d, we draw for it an intent label l_{ds}:

$$l_{ds} \sim \text{Multi}(\text{softmax}(\text{MLP}(h_{ds}))) \tag{5}$$

where h_{ds} is the representation of the current sentence which will be transformed to a probability vector over intent labels via a multi-layer feedforward network. As in the generative process, intent label of a sentence is generated first in Step 2(b), then its words are generated depending on their word types and topics (Step 2(c)). Therefore, h_{ds} of the current sentence needs to be determined based on the previous generated sentences before its words can be generated.

To determine h_{ds} based on previous sentences, as in [11], we append to the end of each sentence a BERT's delimiter token called [SEP] and insert the standard [CLS] token at the beginning of the sentence sequence. Different from [11], as shown in Figure 2, the sentence sequence up to sentence $s-1$ separated by [SEP] is then fed into BERT to determine h_{ds} for predicting the intent label of the current sentence s. Let H_{SEP_s} be the learned representation vector of the [SEP] token associated with the sentence s. As a simple approach, we can let h_{ds} in Eq. 5 be $H_{SEP_{s-1}}$, i.e., the intent label of the current sentence will be determined based on the representation of the previous sentence. This is a reasonable choice because sentences with the same intent are usually put next to each other [14]. However, this approach does not explicitly utilize the topic information in previous sentences for generating the current sentence's intent. In the next section, we introduce a topic attention mechanism which allows the sentence sequence model to attend to topics of previous relevant sentences.

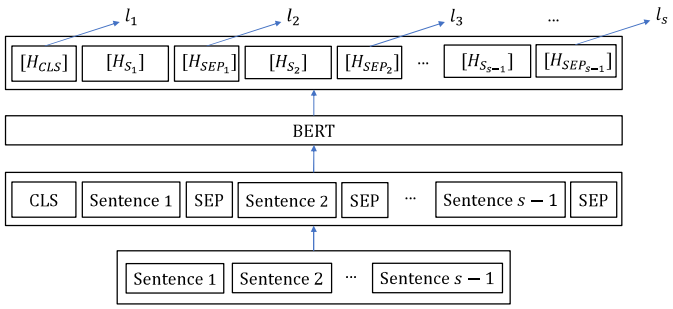


Fig. 2: Using BERT to learn the contextualized representations of previous sentences to predict the intent label of the current sentence. For each sentence s, H_{SEP_s} is the learned contextualized representation vector of the [SEP] token associated with that sentence.

C. Topic Attention Model

Since the above sentence sequence model is integrated with the neural topic model, we can utilize topic words to feed additional information into BERT model to improve further the prediction performance. More specifically, we propose a topic attention model to enhance the performance of document structure learning by enabling the model to attend to the most relevant topic features from all previous sentences. For each sentence s in document d, let C_{ds} be the attention over topic features of all previous sentences from 1 to $s-1$:

$$C_{ds} = \sum_{u=1}^{s-1} \alpha_{du} T^+_{du} \tag{6}$$

When generating sentence s, all previous sentences have been generated. Therefore, we know the topic assignment of each word in the previous sentences. Based on that, we define T^+_{du} to be the sum of embeddings of topics assigned to words within sentence du, $T^+_{du} = \sum_{dum} T_{w_{dum}}$. Since topic assignment is soft, $T_{w_{dum}}$ or the topic embedding of word w_{dum} is calculated as a weighted sum of topic embeddings:

$$T_{w_{dum}} = \sum_{k=1}^{K} p(k|w_{dum}, d) T_k \tag{7}$$

where $p(k|w_{dum}, d)$ is the probability that topic k is assigned to word w_{dum} in document d. $p(k|w_{dum}, d)$ is computed as:

$$p(k|w_{dum}, d) \sim p(w_{dum}|k, d) p(k|d) \tag{8}$$

For the attention mechanism, the attention score of sentence du is given as:

$$a_{du} = (T^+_{du})^\top U \tag{9}$$

where U is the context vector that can be interpreted as a query asking for the most informative sentence from the sentence sequence [17]. U will be learned in the inference algorithm. Based on that, the attention weight α_{du} in Eq. 6 is:

$$\alpha_{du} = \frac{\exp(a_{du})}{\sum_{u'=1}^{s-1} \exp(a_{du'})} \tag{10}$$

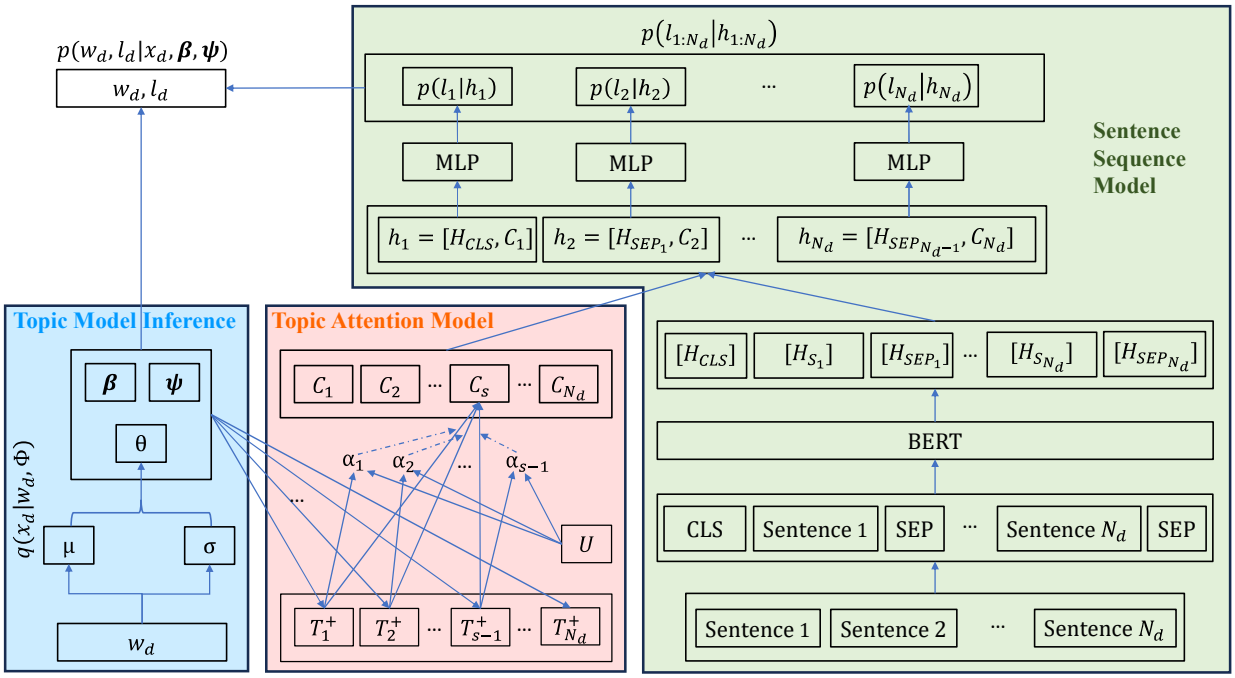


Fig. 3: The overall inference architecture of INTENT.

We then concatenate C_{ds} and $H_{SEP_{s-1}}$ to form the representation of a sentence, h_{ds}, in Eq. 5:

$$h_{ds} = \left[H_{SEP_{s-1}}, C_{ds}\right] \tag{11}$$

This approach defines our method INTENT where we integrate a sentence sequence model and a neural topic model via the topic attention model.

D. Variational Inference

We derive a neural variational inference algorithm to infer the model parameters [18]. Given a document w_d and its sequence of sentence labels l_d, by collapsing b, k, we have the following ELBO:

$$\begin{aligned} L = E_{q(x_d|w_d,\Phi)}\left[\log p(w_d, l_d|x_d, \boldsymbol{\beta}, \boldsymbol{\psi})\right. \\ \left. -D_{KL}(q(x_d|w_d, \Phi)||p(x_d))\right] \end{aligned} \tag{12}$$

where,

$$\begin{aligned} &\log p(w_d, l_d|x_d, \boldsymbol{\beta}, \boldsymbol{\psi}) \\ &= \sum_{s=1}^{N_d} \log p(l_{ds}|h_{ds}) + \sum_{s=1}^{N_d} \sum_{dsm} \log p(w_{dsm}|l_{ds}, \boldsymbol{\beta}, \boldsymbol{\psi}, x_d) \end{aligned} \tag{13}$$

$$\begin{aligned} &p(w_{dsm}|l_{ds}, \boldsymbol{\beta}, \boldsymbol{\psi}, x_d) = \\ &\sum_k p(w_{dsm}|k, \boldsymbol{\psi}) p(k|x_d) p(b=0) + p(w_{dsm}|\boldsymbol{\beta}, l_{ds}) p(b=1) \end{aligned} \tag{14}$$

We maximize the ELBO to learn model parameters. To approximate the expectation in Eq. 12, $q(x_d|w_d, \Phi)$ which is the variational approximation of the posterior will be parameterized by an inference network with Φ be the variational parameters. For the inference network, we use two linear layers to transform the bag-of-words vector w_d to variational Gaussian parameters μ_d, σ_d^2 and use the reparameterization trick to sample $x_d = \mu_d + \sigma_d \epsilon$, with $\epsilon \sim \mathcal{N}(0, 1)$ [18]. θ_d is then computed as $\text{softmax}(x_d)$ (as illustrated in the Topic Model Inference block in Figure 3 and lines 3, 4 in Algorithm 1). In the next step, for each sentence s in document d, we compute the attention over topic features of all previous sentences, C_{ds}, using Eq. 6. The details are depicted in the Topic Attention Model block in Figure 3 and line 6 in Algorithm 1. The sentence sequence model block in Figure 3 summarizes how topic attention model is integrated with the sentence sequence model for inference. More specifically, we compute $h_{ds} = \left[H_{SEP_{s-1}}, C_{ds}\right]$ and then the log likelihood as in Eq. 13 (lines 7-10 in Algorithm 1), where $p(l_{ds}|h_{ds})$ is computed by using a softmax activation on the output of the multi-layer feedforward network $\text{MLP}(h_{ds})$. We optimize the ELBO to learn parameters using Adam optimizer [19]. The overall architecture of the inference network is shown in Figure 3 and the inference algorithm of our proposed model is summarized in Algorithm 1.

III. Experiments

A. Datasets and Baselines

We conduct extensive experiments to demonstrate the effectiveness of INTENT using four real-world datasets from different domains:

- PubMed20k [20]: consists of 20000 abstracts for medical and biological scientific papers. Each sentence is classi-

Algorithm 1 Inference algorithm of INTENT

Input: Training documents from $\mathcal{D}$ with sentence intent labels; K topics; I intents.
Output: Topic distributions of documents ($\boldsymbol{\theta}$); word distributions of topics ($\boldsymbol{\psi}$); word distributions of intents ($\boldsymbol{\beta}$).

for each epoch **do**
  for each document w_d **do**
    Compute μ_d, σ_d^2 from the inference network
    Draw $\theta_d \sim \mathcal{LN}(\mu_d, \sigma_d^2)$
    for each sentence w_{ds} **do**
      Compute the attention of topic features C_{ds} (Eq. 6)
      Obtain $H_{SEP_{ds-1}}$ from sentence w_{ds-1}
      Compute $h_{ds} = \left[H_{SEP_{s-1}}, C_{ds}\right]$
    end for
    Compute $p(w_d, l_d | x_d, \boldsymbol{\beta}, \boldsymbol{\psi})$ by (Eq. 13)
  end for
  Compute the ELBO loss in (Eq. 12)
  Update parameters using Adam optimizer [19]
end for

fied into one of five intent labels: Background, Objective, Methods, Results, and Conclusions.

- NICTA-PIBOSO [21]: consists of 1000 biomedical abstracts with 6 intent labels: Background, Other, Intervention, Study Design, Population, and Outcome.
- Chemical [22]: contains 965 chemical abstracts with 7 intent labels: Background, Objective, Related Work, Method, Result, Conclusion, and Future Work.
- CSAbstract [23]: consists of 654 abstracts collected and annotated from arXiv focusing on applied computer science for social media. Each sentence is categorized into one of five intent labels: Background, Objective, Methods, Results, and Conclusions.

We compare our proposed model with several state-of-the-art methods for supervised document structure learning:

- Supervised topic models:
  - sEGMM-LDA [14]: A supervised topic model for document structure learning. It adopts the generalized Mallows model (GMM) prior to learn the rhetorical order of sentences within a document.
  - sLDA [24]: It is a supervised topic model where each input document is associated with a label. It does not aim to extract intents of sentences.
  - sDTM [25]: It is a supervised neural topic model. By integrating RNN and topical attention mechanism, it is able to show improvement of document classification tasks and document modeling.
- Sequential sentence classification models:
  - HSLN [13][2]: It presents a hierarchical LSTM/CNN + CRF network for sequential sentence classification. There are two variants of the HSLN model: HSLN-CNN and HSLN-RNN.
  - SciBERT [11][3]: A BERT-based model for sequential sentence classification task. It fine-tunes the pre-trained weights of a BERT-based model trained on scientific texts and uses the representation vector of the [SEP] token at the end of the sentence to predict the sentence intent label.
  - SciBERT-HSLN, MULT [10][4]: In this approach, they use HSLN architecture for classification with pre-trained word embeddings from SciBERT. MULT is designed for multi-task learning, which can generally improve performance and has a better generalization, MULT uses the same architecture as SciBERT-HSLN, however, it simultaneously trains samples in all tasks and except for the output layers, all other layers are shared across the tasks.
- INTENT[5]: our proposed model integrates a sentence sequence model with a neural topic model for learning intent and topic structures of documents. In the experiments, for the pre-trained BERT-based language model, we use SciBERT and fine-tune it for the document structure learning task.

We preprocess the data by removing all stop words and keep the top 5000 most frequent words. For all datasets, we split them into train/dev/test sets with the ratio of 60%/20%/20% respectively. In our experiments, all testing results on the test set are reported using the model having the best performance (micro F1 score) on the dev set for sequential classification task. For supervised topic models such as sLDA and sDTM, we split all the sentences within a document and treat each sentence as a single input document with its label. For all datasets, we set the hidden size of all linear layers to 250, learning rate of topic model is searched in $\{0.0005, 0.001, 0.005\}$ and the learning rate of SciBERT is 1e-5, the number of batch size is searched in $\{16, 18, 20, 22\}$. For the optimizer, we use Adam optimizer with weight decay 0.001. For all baselines, we choose the values of hyper-parameters based on the suggestions in their papers. All experimental results are obtained by running 100 epochs and averaged across 4 independent runs.

B. Sequential Sentence Classification

As stated above, identifying intents of sentences in a document can be formulated as a sequential sentence classification task. Effective methods for document structure learning are expected to perform well in this task. In this section, we report the micro F1 score by several supervised methods for predicting sentence intent labels. In Figure 4, it becomes evident that our proposed model INTENT significantly outperforms other baselines in many cases according to the two-sided paired t-test. INTENT has the highest micro F1 score across different numbers of topics and datasets. Supervised topic models such as sEGMM-LDA, sLDA, and sDTM have the lowest micro F1 score because they do not model the sequential context in sentence sequence. The significant gap between our model and the supervised topic models also shows the benefits of

[2]https://github.com/jind11/HSLN-Joint-Sentence-Classification
[3]https://github.com/allenai/sequential_sentence_classification
[4]https://github.com/TIBHannover/sequential-sentence-classification-extended
[5]https://github.com/dangpnh2/INTENT

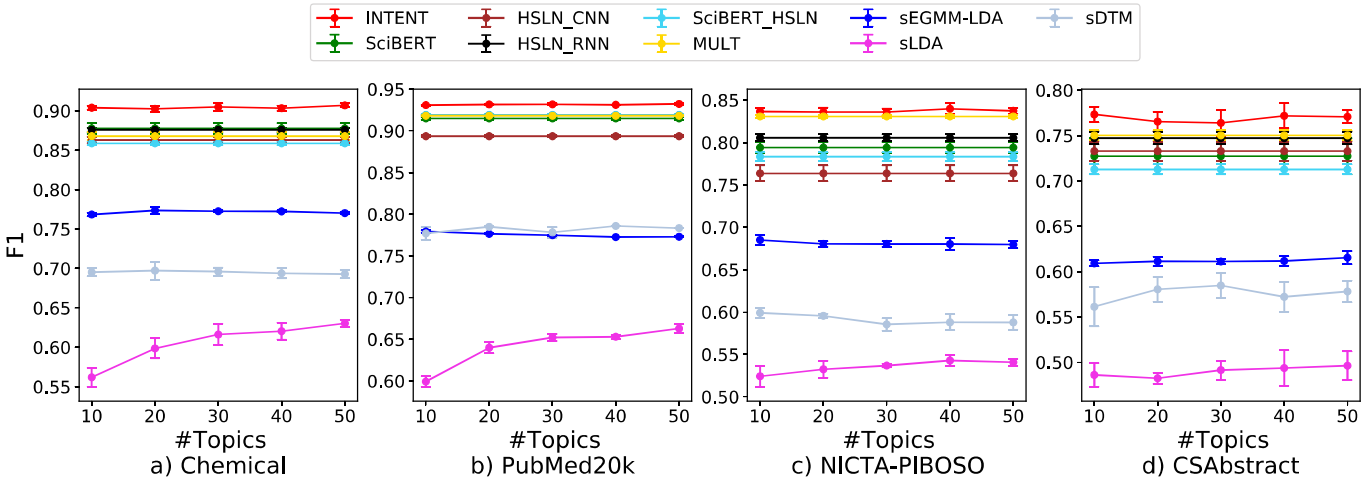


Fig. 4: Micro F1 score of all methods for sequential sentence classification task with different number of topics K.

TABLE I: Micro F1 score of all methods for number of topics $K = \{10, 50\}$. ▲ ($p < 0.005$) and △ ($p < 0.05$) are cases where INTENT is significantly better than the baseline w.r.t. the two-sided paired t-test.

Category	Model	Chemical		PubMed20k		NICTA-PIBOSO		CSAbstract	
		10	50	10	50	10	50	10	50
Supervised topic models	sLDA	56.19±1.21▲	63.04±0.42▲	59.93±0.63▲	66.29±0.50▲	52.42±1.25▲	54.07±0.42▲	48.61±1.32▲	49.62±1.58▲
	sDTM	69.53±0.52▲	69.29±0.52▲	77.70±0.74▲	78.34±0.00▲	59.91±0.59▲	58.78±0.85▲	56.12±2.17▲	57.81±1.20▲
	sEGMM-LDA	76.85±0.23▲	77.03±0.17▲	77.94±0.12▲	77.31±0.06▲	68.50±0.61▲	67.96±0.43▲	60.91±0.33▲	61.54±0.71▲
Sequential sentence classification models	HSLN_CNN	86.30±0.40▲	86.30±0.40▲	89.37±0.10▲	89.37±0.10▲	76.35±0.93▲	76.35±0.93▲	73.28±1.04△	73.28±1.04△
	HSLN_RNN	87.59±0.26▲	87.59±0.26▲	91.87±0.08▲	91.87±0.08▲	80.55±0.40▲	80.55±0.40▲	74.72±0.69△	74.72±0.69△
	SciBERT_HSLN	85.88±0.23▲	85.88±0.23▲	91.92±0.07▲	91.92±0.07▲	78.33±0.54▲	78.33±0.54▲	71.27±0.56▲	71.27±0.56▲
	MULT	86.97±0.39▲	86.97±0.39▲	91.89±0.06▲	91.89±0.06▲	83.09±0.59	83.09±0.59△	75.03±0.55△	75.03±0.55△
	SciBERT	87.76±0.67△	87.76±0.67▲	91.50±0.15▲	91.50±0.15▲	79.41±0.60▲	79.41±0.60▲	72.73±1.98△	72.73±1.98△
Joint supervised	INTENT	**90.40±0.25**	**90.69±0.26**	**93.10±0.06**	**93.23±0.09**	**83.42±0.23**	**83.55±0.48**	**77.33±0.86**	**77.07±0.72**

fine-tuning a pre-trained language model such as SciBERT for modeling sentence sequence in document structure learning.

When comparing among the sequential sentence classification methods, MULT is the best baseline in terms of micro F1 score. This approach shows a significant performance gap to its base model SciBERT-HSLN. This shows that all these models benefit from fine-tuning pre-trained language models for the sequential sentence classification task. Since our model integrates topic modeling and intent prediction via the topic attention model, it can surpass these strong baselines. Table I shows the detailed micro F1 score of all methods for different number of topics $K = \{10, 50\}$. For all settings, INTENT is the method that has the highest micro F1 score. INTENT is significantly better than other baselines in most cases according to the two-sided paired t-test.

C. Document Modeling

We show that while jointly extracting topics and predicting intents, INTENT is still an effective supervised topic model and can be generalized to model unseen documents. We rely on the document modeling task which aims to measure the generalization performance of topic models. Topic models that are good at document modeling would achieve high likelihood on a held-out test set. We compute the perplexity of a held-out test set to evaluate the models:

$$\text{perplexity}(\text{D}_{\text{test}}) = \exp\left[-\frac{\sum_d \log p(w_d)}{\sum_d M_d}\right] \qquad (15)$$

where $\log p(w_d)$ is the log likelihood, and M_d is the number of tokens in the held-out test set [15]. Figure 5 shows the test perplexity by different topic models across different numbers of topics K. In this Figure, INTENT consistently achieves lower perplexity scores compared to the baseline methods across all three datasets and numbers of topics. For example, on the PubMed20k dataset, INTENT demonstrates significantly lower perplexity values as the number of topics increases, indicating its ability to model the underlying distribution of the data. Specifically, while other models exhibit a rapid increase in perplexity as the number of topics increases, INTENT maintains a more stable and lower perplexity curve, underscoring its robustness and efficiency in handling complex text data. These findings show that our models have good generalization performance while also performing well in intent prediction.

D. Topic Coherence

Topic coherence is widely used to evaluate the quality of generated topics. We use Normalize Pointwise Mutual Information (NPMI), a widely used metric for topic coherence measurement. In our experiments, we estimate the pair score by using a large external Wikipedia dataset. Figure 6 shows the average NPMI score over all topics of each method. In this figure, INTENT, sEGMM-LDA are the methods that have the highest score over all datasets, except for NICTA-PIBOSO, INTENT outperforms other models in all number of topics. While sDTM and sLDA show lower NPMI values, indicating weaker topic coherence. In general, INTENT and sEGMM-LDA show a comparable results. The results show that our proposed model can extract coherent topics. Some examples of topic words extracted by INTENT are shown in Figure 8.

E. Example of Intent and Topic Words

In our model, each word in a document is either a topic word or an intent word. To see how well our model separates these two word types, we show some top word examples in Figures 7,8, and Figure 9a. Intent words convey the underlying purpose or objective of the sentence and guide the structure of a document. Each intent word tends to appear in some specific section within a document. In Figure 7, we visualize top intent words generated by INTENT on PubMed20k dataset using word clouds with weights as word probabilities. We can see that these words aptly capture the intent labels. For instance, in Figure 7a, the top intent words are "compare", "evaluate", "determine", "investigate", and "aim" which indicate clearly the "Objective" intent label. Figure 7e shows representative words such as "randomized", "trial", "using", "controlled" for the "Method" intent label, and Figure 7c shows the top words "effective", "results", "finding", "suggest", "improve" which represent reasonably the "Conclusion" intent label.

Topic words, on the other hand, provide the subject matter or the theme being discussed in a document. They are more likely to be distributed across the documents in the corpus. Other than intent words, these words convey or indicate the domain of documents. Figure 8 shows the word clouds of ten topics belonging to PubMed20k dataset. For example, "group", "control", "treatment", "effect", and "patients" in Topic 0 represent clinical trials comparing control and treatment groups of patients. Topic 2 is closely related to clinical trials that compare drug treatments with placebos, indicated through terms such as "placebo", "dose", "patients", "response", and "mg". Top words such as "exercise", "training", "sleep", "physical", and "depression" in Topic 5 show the study on physical activities, treatments to improve mental and physical healh problems. The last topic focuses on "mortality", "risk", "cardiac", "blood", and "pressure" which discuss the risk factors and mortality rates associated with cardiac events like heart failure and stroke.

We also highlight a sample document from PubMed20k dataset in Figure 9a. As we can see, intent words and topic words are well splitted. Blue words are intent words and red words are topic words. The gray words are either stop words or words with low frequency which are removed in the preprocessing step before training.

F. Topic Attention Example

Figure 9a shows an example abstract from PubMed20k dataset whose corresponding attention map of all sentences is presented in Figure 9b. The attention map shows how each sentence (each row) attends to the topic embeddings of all previous sentences. The attention weights are obtained from Eq. 6. As we can see, sentence S2 puts full attention weight on sentence S1 as S1 is the only sentence before S2. Sentences S3, S4 depend on the topic embeddings of sentence S2 whose topics are (2) and (7). This could be explained by the fact that S3, S4 share similar topics with S2 and hence they tend to have the same intent label. Although S5 shares similar topics with S2, S5 and S2 do not necessarily have the same intent label. This shows that the model also needs to rely on intent words to predict intents correctly. As we can see, S5 contains some intent words such as "eleven percent" and "compared" which clearly indicates its "RESULTS" intent. For sentences S6-S10, the model tends to attend to topics (1) and (2) in S5 to explain the intents of S6-S10, which is reasonable because S6-S10 also contain both topics (1) and (2) as in S5. This shows that adjacent sentences with similar topics often have the same intents. For S11-S12, a combination of topic words such as "groups", "data", "study" and intent words such as "but", "support", "hypothesis", "could prevent" may help to predict correctly their "CONCLUSIONS" intent label. Typically, a mention of these words such as "support", "hypothesis" in sentences right after the "RESULTS" sentences likely implies that the writer wants to conclude something about the current investigated hypothesis. In this case, the considered hypothesis in the example abstract is not supported by the data.

IV. Related Work

A. Supervised Topic Models

Supervised topic models integrate labeled information into the topic modeling process, enabling the incorporation of domain-specific knowledge or supervision signals [24], [26], [27]. They have been applied to a wide range of applications such as document classification [28], [29], sentiment analysis [30], [31], and regression [32] One of the most widely used methods is sLDA which extends LDA by incorporating the real value of the document label to guide the topic modeling process [24]. By leveraging labeled data, supervised topic models can learn topics that are more aligned with the specific tasks or domains of interest. In [33] and [25], the authors propose models to enhance the performance of document classification task by utilizing the recurrent neural network and the attention mechanism to capture the word order. None of these methods model intents at sentence level as well as the sequence order of sentences in the document for document structure learning.

B. Sequential Sentence Classification

In recent years, deep learning approaches have shown several advantages for text classification tasks, as they are

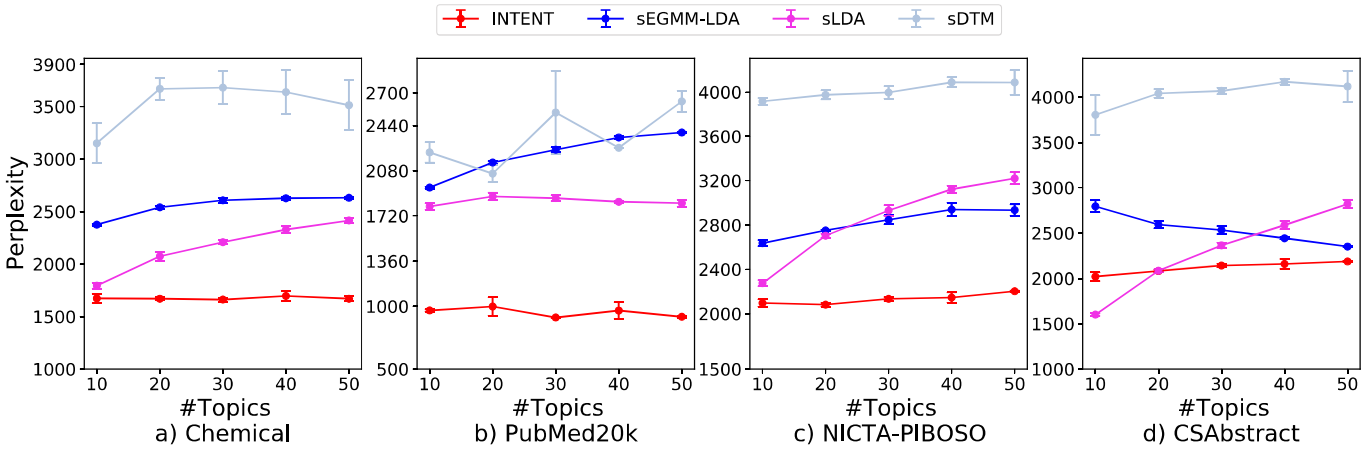


Fig. 5: Perplexity on the held-out test set by topic modeling methods (lower is better).

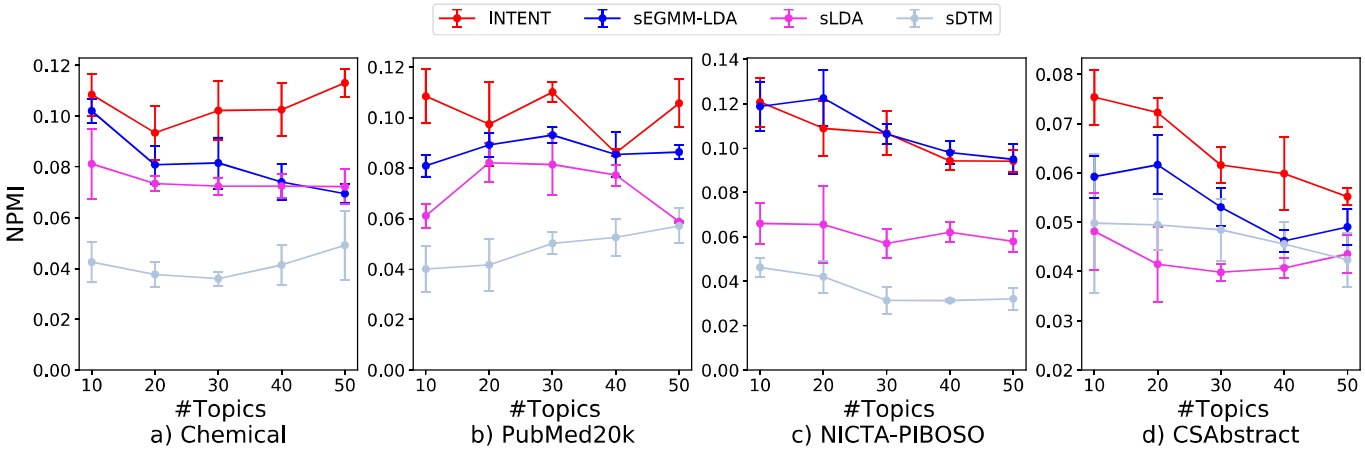


Fig. 6: Topic coherence (NPMI score) by topic modeling methods (higher is better).

(a) Objective (b) Background (c) Conclusions (d) Results (e) Methods

Fig. 7: Word clouds showing the top words of five intents generated by INTENT on PubMed20k.

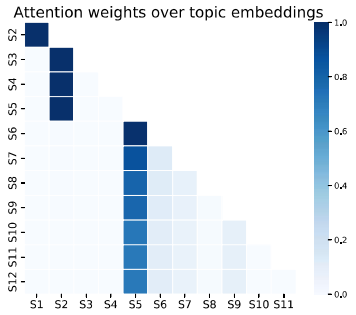
(a) Topic 0 (b) Topic 1 (c) Topic 2 (d) Topic 3 (e) Topic 4

(f) Topic 5 (g) Topic 6 (h) Topic 7 (i) Topic 8 (j) Topic 9

Fig. 8: Word clouds showing the top words of ten topics generated by INTENT on PubMed20k.

(S1) OBJECTIVE: The purpose of this study(2) was to determine whether @ g(3) kg(3) oral(2) ivermectin is able to kill Ixodes scapularis nymphs and adult(2) female(7) ticks feeding(7) on humans(3)
(S2) METHODS: Ten study(2) subjects(2) each wore @ ostomy bags , the one containing @ I scapularis nymphs , and the other containing @ I scapularis adult(2) females(7)
(S3) METHODS: Twenty four hours(1) after the ostomy bags were attached , study(2) subjects(2) received either @ g(3) kg(3) ivermectin or placebo(2)
(S4) METHODS: Thirty hours(1) after the ivermectin or placebo(2) was consumed , the ticks were removed , and mortality(9) determined in a double-blinded manner .
(S5) RESULTS: Eleven percent of the I scapularis nymphs attached in the ivermectin group(1) compared with @ % in the placebo(2) .
(S6) RESULTS: Mortality(9) for the I scapularis nymphs that attached at the time(1) of removal was @ % in the ivermectin group(1) and @ % in the placebo(2) group(1) .
(S7) RESULTS: Mortality(9) for the I scapularis nymphs @ days(2) after removal was @ % in the ivermectin group(1) and @ % for the placebo(2)
(S8) RESULTS: Three percent of the I scapularis adults(2) attached in the ivermectin group(1) compared with @ % in the placebo(2) group(1) .
(S9) RESULTS: Mortality(9) for I scapularis adults(2) was @ % on day @ and @ % on day(2) @ for both the ivermectin and placebo(2) groups(1)
(S10) RESULTS: There were statistically insignificant differences in the mortality(9) rates between I scapularis nymphs and adults(2) exposed to ivermectin or placebo(2) .
(S11) CONCLUSIONS: There were a high number of ticks that died in both groups(1) but the data(2) do not support our hypothesis that ivermectin can kill I scapularis .
(S12) CONCLUSIONS: The study(2) was not designed to determine whether it could prevent the transmission of tick-borne illness(8) .

(a) Word type for each word in an abstract from PubMed20k dataset (blue: intent word, red: topic word) and their topic assignments (topic number is put next to each topic word).



(b) Attention weights.

Fig. 9: Example attention weights over topic embeddings.

able to learn meaningful representations for texts. HSLN [13] proposes a hierarchical neural network followed by a CRF layer. It employs bi-LSTM/CNN as a word/sentence encoder to capture the sequential dependencies between words, and sentences. The work in [34] introduces Neural Semi-Markov Conditional Random Fields SCRFs to handle this problem at span level (segment) instead of sentence level. More recently, large language models (LLMs) such as GPT and BERT have shown a significant improvement in understanding context and capturing more relevant features by leveraging vast amounts of text data, leading to significant improved performance in various sentence classification tasks. The work in [11] utilizes SciBERT [35] which is a BERT-based model trained on roughly 1.14M scientific documents and performs fine-tuning for the sequential sentence classification task. As another approach, SciBERT-HSLN employs transfer learning for sequential classification tasks within the scientific domain including sequential transfer learning (INIT) and multi-task learning (MULT) [10]. In INIT, they first train the model on source data and then update the parameters using the different target data. In contrast, MULT trains all the tasks simultaneously and only the output layers are different for all tasks while the remaining parameters are shared across the tasks. Most of these methods focus on utilizing rhetorical words but ignore topics of sentences.

C. Document Structure Learning

Learning document structure is crucial for many NLP tasks, as it provides a deeper understanding of the relationships between different parts of a document (could be at sentence or segment level). Traditional models often treat documents as a bag of words, ignoring the inherent structure that can provide valuable context. Recent approaches have sought to overcome this limitation by explicitly modeling the hierarchical structure of documents. One such method involves using hierarchical models that capture the relationships between sentences and paragraphs, providing a deeper comprehension of document composition. For example, the Hierarchical Attention Network (HAN) introduced by [36] uses hierarchical attention mechanisms to focus on important words within sentences and important sentences within documents. This allows the model to capture both the local (word-level) and global (sentence-level) context. Tree-LSTM uses tree-structured neural networks to represent the syntactic structure of text, which allows to learn a more natural and effective representation of the hierarchical structure of text [37].

Recently, the integration of rhetorical structure theory (RST) has been explored to enhance document structure learning. RST focuses on functional relationships between different parts of a text, such as contrast, elaboration, and cause-effect. Different approaches have been introduced, for instance, the work in [38] develops a discourse parser that combines RST with neural networks, achieving state-of-the-art performance in discourse parsing. Another work in [39] incorporates graph theory and RST for relation extraction at the document level. [40] is another work that utilizes graph and attention network for relation extraction problem. Lastly, a closely related work to our model is sEGMM-LDA [14]. sEGMM-LDA incorporates RST into topic models and distinguishes between topical and rhetorical words, which improves the understanding of document coherent structure. Different from INTENT, sEGMM-LDA does not model intents at sentence level as well as the sequence nature of words and sentences in the document, which makes it not perform well in the sequential sentence classification task.

V. Conclusion

In this paper, we propose a model that integrates a sentence sequence model with a neural topic model for supervised learning of document structure which can be captured through topics and intents of sentences in the document. We design a topic attention mechanism that utilizes the topics of previous sentences for improving the intent inference performance. Extensive experiments on four real-world datasets show the effectiveness of our proposed model. This validates the joint modeling approach of topics and intents by integrating a language model for sequence modeling with a topic model for extracting topics via the topic attention mechanism.

ACKNOWLEDGMENT

This research is partially sponsored by NSF award #1914635.

REFERENCES

[1] X. Jin and Y. Wang, "Understand legal documents with contextualized large language models," *arXiv preprint arXiv:2303.12135*, 2023.

[2] S. R. Ahmad, D. Harris, and I. Sahibzada, "Understanding legal documents: classification of rhetorical role of sentences using deep learning and natural language processing," in *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*. IEEE, 2020, pp. 464–467.

[3] T. S. Santosh, P. Bock, and M. Grabmair, "Joint span segmentation and rhetorical role labeling with data augmentation for legal documents," in *European Conference on Information Retrieval*. Springer, 2023, pp. 627–636.

[4] V. Malik, R. Sanjay, S. K. Guha, A. Hazarika, S. Nigam, A. Bhattacharya, and A. Modi, "Semantic segmentation of legal documents via rhetorical roles," *arXiv preprint arXiv:2112.01836*, 2021.

[5] P. Accuosto, M. Neves, and H. Saggion, "Argumentation mining in scientific literature: From computational linguistics to biomedicine," in *Frommholz I, Mayr P, Cabanac G, Verberne S, editors. BIR 2021: 11th International Workshop on Bibliometric-enhanced Information Retrieval; 2021 Apr 1; Lucca, Italy. Aachen: CEUR; 2021. p. 20-36*. CEUR Workshop Proceedings, 2021.

[6] P. Accuosto and H. Saggion, "Mining arguments in scientific abstracts with discourse-level embeddings," *Data & Knowledge Engineering*, vol. 129, p. 101840, 2020.

[7] S. A. Takale, "Knowledge representation for legal document summarization," *International Journal of Innovative Research in Computer Science & Technology*, vol. 11, no. 4, pp. 61–66, 2023.

[8] T. Goldsack, Z. Zhang, C. Lin, and C. Scarton, "Domain-driven and discourse-guided scientific summarisation," in *European Conference on Information Retrieval*. Springer, 2023, pp. 361–376.

[9] J. Atkinson and R. Munoz, "Rhetorics-based multi-document summarization," *Expert Systems with Applications*, vol. 40, no. 11, pp. 4346–4352, 2013.

[10] A. Brack, E. Entrup, M. Stamatakis, P. Buschermöhle, A. Hoppe, and R. Ewerth, "Sequential sentence classification in research papers using cross-domain multi-task learning," *International Journal on Digital Libraries*, pp. 1–24, 2024.

[11] A. Cohan, I. Beltagy, D. King, B. Dalvi, and D. S. Weld, "Pretrained language models for sequential sentence classification," *arXiv preprint arXiv:1909.04054*, 2019.

[12] Y. S. S. S. Tokala, S. S. Aluru, A. Vallabhajosyula, D. K. Sanyal, and P. P. Das, "Label informed hierarchical transformers for sequential sentence classification in scientific abstracts," *Expert Systems*, vol. 40, no. 6, p. e13238, 2023.

[13] D. Jin and P. Szolovits, "Hierarchical neural networks for sequential sentence classification in medical scientific abstracts," *arXiv preprint arXiv:1808.06161*, 2018.

[14] B. Chen, J. Zhu, N. Yang, T. Tian, M. Zhou, and B. Zhang, "Jointly modeling topics and intents with global order structure," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[16] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[17] X. Sun and W. Lu, "Understanding attention for text classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3418–3428.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] F. Dernoncourt and J. Y. Lee, "Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts," *arXiv preprint arXiv:1710.06071*, 2017.

[21] S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken, "Automatic classification of sentences to support evidence based medicine," in *BMC bioinformatics*, vol. 12, no. 2. BioMed Central, 2011, pp. 1–10.

[22] Y. Guo, A. Korhonen, M. Liakata, I. Silins, L. Sun, and U. Stenius, "Identifying the information structure of scientific abstracts: an investigation of three different schemes," in *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, 2010, pp. 99–107.

[23] S. Gonçalves, P. Cortez, and S. Moro, "A deep learning classifier for sentence classification in biomedical and computer science abstracts," *Neural Computing and Applications*, vol. 32, no. 11, pp. 6793–6807, 2020.

[24] J. Mcauliffe and D. Blei, "Supervised topic models," *Advances in neural information processing systems*, vol. 20, 2007.

[25] Y. Yang, K. Zhang, and Y. Fan, "sdtm: A supervised bayesian deep topic model for text analytics," *Information Systems Research*, vol. 34, no. 1, pp. 137–156, 2023.

[26] J. Zhu, A. Ahmed, and E. P. Xing, "Medlda: maximum margin supervised topic models for regression and classification," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 1257–1264.

[27] D. Sridhar, H. Daumé III, and D. Blei, "Heterogeneous supervised topic models," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 732–745, 2022.

[28] F. Rodrigues, M. Lourenco, B. Ribeiro, and F. C. Pereira, "Learning supervised topic models for classification and regression from crowds," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2409–2422, 2017.

[29] X. Li, J. Ouyang, and X. Zhou, "Supervised topic models for multi-label classification," *Neurocomputing*, vol. 149, pp. 811–819, 2015.

[30] J. Pang, Y. Rao, H. Xie, X. Wang, F. L. Wang, T.-L. Wong, and Q. Li, "Fast supervised topic models for short text emotion detection," *IEEE Transactions on Cybernetics*, vol. 51, no. 2, pp. 815–828, 2019.

[31] F. Li, S. Wang, S. Liu, and M. Zhang, "Suit: A supervised user-item based topic model for sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, no. 1, 2014.

[32] M. Magnusson, L. Jonsson, and M. Villani, "Dolda: a regularized supervised topic model for high-dimensional multi-class regression," *Computational Statistics*, vol. 35, no. 1, pp. 175–201, 2020.

[33] X. Wang and Y. Yang, "Neural topic model with attention for supervised learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1147–1156.

[34] K. Yamada, T. Hirao, R. Sasano, K. Takeda, and M. Nagata, "Sequential span classification with neural semi-markov crfs for biomedical abstracts," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 871–877.

[35] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[36] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[37] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[38] Q. Li, T. Li, and B. Chang, "Discourse parsing with attention-based hierarchical neural networks," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 362–371.

[39] H. Wang, K. Qin, G. Lu, J. Yin, R. Y. Zakari, and J. W. Owusu, "Document-level relation extraction using evidence reasoning on rst-graph," *Knowledge-Based Systems*, vol. 228, p. 107274, 2021.

[40] Q. Sun, K. Zhang, K. Huang, T. Xu, X. Li, and Y. Liu, "Document-level relation extraction with two-stage dynamic graph attention networks," *Knowledge-Based Systems*, vol. 267, p. 110428, 2023.