Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.087209

ARTICLE

Enhancing Biomedical Multi-Label Text Classification via Topic-Based Text Representation

Oyku Berfin Mercan[1,2], **Nezihe Turhan Turan**[3] **and Aytuğ Onan**[4,*]

[1]R&D Department, Türk Telekom, Ankara, Türkiye
[2]Department of Computer Engineering, İzmir Katip Celebi University, İzmir, Türkiye
[3]Department of Engineering Sciences, İzmir Katip Celebi University, İzmir, Türkiye
[4]Department of Computer Engineering, İzmir Institute of Technology, İzmir, Türkiye

*Corresponding Author: Aytuğ Onan. Email: aytugonan@iyte.edu.tr

Received: 12 June 2026; Accepted: 07 August 2026; Published: 15 September 2026

ABSTRACT: Biomedical texts naturally contain multiple biological and medical concepts within a document, resulting in a semantically rich and complex structure. Consequently, multi-label text classification (MLTC) has become a suitable framework for comprehensively modeling biomedical texts, including clinical reports, laboratory records, and scientific abstracts. However, relying solely on contextual language representations may be insufficient to explicitly reflect the broader scientific focus and conceptual orientation of a document. In this study, the MLTC problem in the biomedical domain is investigated using the Hallmarks of Cancer (HoC) dataset. Topic probability distributions obtained from CombinedTM are incorporated as an additional representational signal into pre-trained language models (PLMs), enabling the joint exploitation of contextual semantic information and document-level topical characteristics. Both general-purpose and biomedical language models were fine-tuned and evaluated within this unified framework. The experimental results demonstrate that incorporating topic information substantially improves the domain robustness of the models. In particular, the Macro-F1 score of the general-purpose BERT model increases from 0.7604 to 0.8458, indicating that it becomes competitive with biomedical language models on tasks that require biomedical domain knowledge. Similarly, biomedical models such as BioMed-RoBERTa also benefit from topic modeling, with Macro-F1 improving from 0.8665 to 0.8819; meanwhile, the highest overall performance is achieved by the PubMedBERT + CombinedTM configuration, with a Macro-F1 score of 0.8900. These findings indicate that integrating topic modeling–based thematic information with PLMs provides an effective, generalizable, and practical solution for biomedical MLTC tasks. Moreover, enriching general-purpose language models with topic information offers a promising alternative that reduces reliance on costly, data-intensive domain adaptation.

KEYWORDS: Biomedical multi-label text classification; pre-trained language model; topic modeling; contextualized topic models; CombinedTM; hallmarks of cancer

1 Introduction

Text classification is one of the fundamental tasks in the field of Natural Language Processing (NLP), as it has paved the way for automatic categorization of increasing amounts of text-based data into meaningful categories based on their content, enabling effective handling of large-scale textual information [1]. Recent studies have highlighted the growing importance of NLP applications in areas such as health informatics, clinical decision support systems, and biomedical information management [2–4]. This trend has placed text classification approaches for automatically categorizing medical texts at the center of healthcare and

NLP research. The rapidly increasing volume of medical literature, electronic health records (EHRs), clinical notes, and research abstracts has made manual review of such data excessively time-consuming and costly [4–6]. Text classification research in the healthcare domain has become central to automation in numerous tasks, including the extraction of disease from symptoms, identification of adverse drug reactions, classification of ICD (International Classification of Diseases) codes for clinical coding, and semantic labeling of scientific publications [7–10]. However, medical texts often cannot be fully represented with traditional single-label classification approaches. By their nature, these texts contain multidimensional, conceptually overlapping, and context-dependent information [11,12]. For instance, in cancer classification, an article abstract might be simultaneously associated with multiple labels such as 'sustaining proliferative signaling', 'evading growth suppressors', and 'resisting cell death' [13]. Similarly, diagnostic reports contained in clinical texts can also be associated with multiple ICD codes [14]. Consequently, this necessitates adopting a multi-label classification approach, in which each instance can be associated with multiple labels. Multi-label classification is more effective in capturing the complex semantic structure of biomedical texts, as it models not only the presence of individual categories but also the co-occurrence and interdependence of multiple conceptual labels [11]. Therefore, multi-label classification problems, particularly in conceptually dense fields of biomedical research, require robust representation methods and learning strategies capable of capturing both the semantic depth of language.

Various research studies have focused on advanced representation strategies and learning approaches to enhance the effectiveness of multi-label classification in the medical domain. Traditional machine learning-based models used in multi-label classification treat labels independently or in a chained manner. In this context, methods such as Binary Relevance (BR), Classifier Chains (CC), Label Powerset (LP), and MLkNN, which enable modeling label dependencies at different levels, are widely used [15]. These models typically utilize text representation techniques such as Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), Word2Vec, or FastText to obtain numerical representations of texts [6]. These methods can provide practical solutions for problems with limited data volume or in scenarios requiring low computational cost. However, the complex terminological structure, context-dependent concept relationships, and multi-layered semantic structure of medical texts limit the expressive power of these models [16,17]. In particular, traditional representation techniques used in numerical representations of textual data may be insufficient to capture the contextual meaning and diversity inherent in medical language. Therefore, this makes it difficult for traditional methods to generalize with high performance in the medical field and reveals the need for deeper representations that can capture contextual meaning. The limitations of conventional approaches have led to the development of transformer-based methods that can model the contextual semantic relationships of texts [18]. Transformer-based pre-trained language models (PLMs) can capture the semantic structure of a language more comprehensively by representing words not only as individual representations but also in the context of their relationships with surrounding words. Through the self-attention mechanism of the transformer architecture, these models can effectively learn long-range dependencies and dynamically redefine the meaning of each word based on its context. General-purpose models such as BERT, RoBERTa, and DistilBERT, trained on large-scale text datasets, offer robust and generalizable language representations. However, these models, which are presented as general-purpose based on the content of their training datasets, lack sufficient representation capabilities for domain-specific structures such as medical terminology, biological process terms, and clinical expression patterns [19]. To address this limitation, domain-specific PLMs such as BioBERT, SciBERT, ClinicalBERT, and PubMedBERT have been introduced [20,21]. These models are further pre-trained on large-scale biomedical text collections, enabling them to adapt to the statistical and semantic characteristics of medical language. Compared to

general-purpose models, they demonstrate improved performance on tasks involving conceptual overlap, rare terminology, and label dependencies.

Despite this contextual representational strength, the information contained in scientific texts is often not limited to intra-sentence relations alone. Revealing the latent thematic structures within texts ensures a more meaningful positioning of documents in the knowledge space [22]. Scientific texts often contain complex information that encompasses many interconnected phenomena, terms, and research topics, rather than a single concept. In this regard, topic modeling methods, which reveal conceptual clusters among the terms contained within the text, play an essential complementary role [23]. By analyzing co-occurring terms in texts, topic modeling reinterprets documents based on latent thematic structures. Thereby, it captures shared scientific relationships across texts with similar conceptual content, facilitating the understanding of thematic clusters and research trends. Thus, the text is represented not only by its vocabulary, but also by its dominant scientific themes. Biomedical literature, due to its terminological density and knowledge structure, requires domain expertise, and the frequent simultaneous treatment of multiple biological processes makes it a highly suitable research area for topic modeling techniques. Topic modeling analyzes the complex terminology in biomedical texts, distinguishing texts by themes, and enabling the transformation of scientific knowledge into more understandable, clusterable, and inferentially powerful representations [23,24].

1.1 Problem Statement

MLTC for biomedical texts is crucial for automated analysis of a wide range of complex and multidimensional sources, from clinical reports to research articles [10,25]. Recent studies have yielded significant findings in the analysis of medical texts using various representation strategies and learning approaches [6,10,25,26]. In particular, pretrained language models have proven effective in understanding the contextual features of texts. However, since their decision-making processes rely solely on textual context, they often fall short of capturing how strongly a document relates to its overarching topic.

Topic modeling methods provide an efficient approach for uncovering latent themes embedded in texts. With the advancement of BERT-based topic modeling frameworks, robust topic distributions can be generated that more accurately represent the conceptual structure of texts across various research domains [27,28]. Although the literature demonstrates that topic modeling has been used for thematic discovery in medical texts, studies integrating topic modeling with multi-label classification remain relatively limited.

On the other hand, biomedical literature constitutes a highly challenging data type for classification due to its rapidly expanding terminology, diverse interdisciplinary concepts, and constantly updated knowledge structure. Although domain-adaptive models tend to perform better against these challenges, the broad-scale data requirements, the need for adaptation, computational costs, and sustainability issues make the development process demanding in several respects. In this context, incorporating high-level thematic signals from documents into the classification pipeline can strengthen decision-making mechanisms. By focusing on the significant performance gains of general-purpose language models through reduced reliance on domain-adaptive models, this approach is a candidate for a practical, adaptable solution for limited data. In this regard, the joint use of contextual text representations and topic-model-derived thematic information is a promising research direction for improving the performance of biomedical MLTC.

1.2 Main Contributions

Scientific texts in the biomedical literature often contain multiple pieces of information simultaneously. As a representative example, texts in the cancer literature can be associated with one or more Hallmarks of Cancer categories, reflecting the multi-label nature of these texts. In this context, the present study is

conducted on the Hallmarks of Cancer (HoC) dataset, which serves as a representative example of the multi-label structure of biomedical literature, and the findings are evaluated within this thematic context.

This study aims to enhance the representational power of texts in the MLTC of biomedical literature by going beyond traditional approaches that rely solely on contextual language representations derived from raw text. In this regard, an integrated representation approach is proposed, in which contextual semantic representations from PLMs and topic probability distributions from topic modeling are fused to jointly inform the model's learning process. This integrated representation of the classification is sensitive not only to local-level contextual relationships but also to the position of the document within the knowledge space. Within this scope, the effects of different model components and configurations on classification performance are comprehensively evaluated. The main contributions of the study can be summarized as follows;

- **Topic-enhanced document representations:** Topic probability distribution vectors were obtained for each document using the CombinedTM (Combined Topic Model), to provide an additional representational strength, and they were integrated with the contextual representations generated by the language model under a single learnable structure.
- **Improving domain robustness of pre-trained models:** It has been observed that incorporating topic probability distributions into the model fine-tuning process significantly improves the classification performance of PLMs. In particular, the improvement achieved on BERT, a general-purpose model, demonstrates that it is competitive with domain-adapted models in multi-label classification tasks that require biomedical domain-specific knowledge (Macro-F1: 0.7604 → 0.8458). Similarly, domain-adaptive models also benefited from topic enrichment (BioMed-RoBERTa: 0.8665 → 0.8819 Macro-F1, ClinicalBERT: 0.8066 → 0.8528 Macro-F1), while the highest overall performance was achieved with the PubMedBERT + CombinedTM configuration with 0.8900 Macro-F1. These findings suggest that enriching general-purpose models with topic information offers a promising solution that may reduce the need for domain-adaptation processes.
- **Comparative evaluation of CombinedTM with various embedding models:** General-purpose and biomedical embedding models were utilized within the contextual representation component of CombinedTM and were evaluated in a comparative manner. The impact of the topic distributions generated by CombinedTM using different models on MLTC performance was analyzed in detail.
- **Benchmarking on the HoC dataset:** The proposed approach was systematically evaluated on the publicly available HoC dataset, and all combinations constructed using different contextual embedding models and PLMs were comparatively analyzed. Through this process, the contribution of topic integration to MLTC was experimentally demonstrated.
- **Evaluation with performance metrics:** Performance analyzes were conducted using metrics suitable for multi-label tasks, including Micro-F1, Macro-F1, Matthews Correlation Coefficient (MCC), Hamming Loss, Jaccard Similarity (Micro/Macro), and Label Ranking Loss. The findings provide practical and methodological guidance for method selection for MLTC by revealing the interaction between data structure and model type.

1.3 Paper Organization

The remainder of this paper is structured as follows: Section 2 presents an overview of existing studies on multi-label biomedical text classification in the literature. Section 3 presents the method proposed in this study. The Section 4 provides details about the experimental setup, including the dataset, models, evaluation metrics, working environment, and model configurations used in the study. Section 5 reports the obtained findings and evaluation results. Section 6 discusses the limitations of the study and research opportunities.

Finally, Section 7 summarizes the main conclusions of the study and discusses potential directions for future research.

2 Related Work

The MLTC literature on medical texts demonstrates significant advances in text representation methods and modeling strategies. Early work in this area focused on feature engineering and traditional machine learning approaches, while subsequent work focused on this direction with the development of transformer-based models. Furthermore, topic modeling approaches play a significant role in medical NLP studies and, thanks to their ability to uncover thematic structures in large-scale text collections, are widely used for tasks such as knowledge discovery and content analysis. This section provides an overview of traditional methods and transformer-based approaches, a popular research topic in recent years, in the context of MLTC studies in the medical field. Then, usage examples of topic modeling will be examined, and its contribution to NLP applications in the biomedical field will be investigated.

2.1 Traditional Approaches for MLTC

The literature reveals the effectiveness of text representation methods and traditional machine learning models in multi-label medical text classification. Multi-label machine learning models that leverage different text representation strategies to model independent or sequential relationships between labels have laid an important foundation for multi-label classification research. In [29], the effectiveness of various machine learning models and CC was investigated for the classification of medical text. The study proposed two ensemble models for text classification, Ensemble of Classifier Chains based on Linear SVC (ECCSVM) and Ensemble of Classifier Chains based on Logistic Regression (ECCLR) both of which demonstrated strong performance. In comparative analyzes, ECCSVM stands out with an F-measure of 0.924 micro, precision of 0.872 micro, and recall of 0.927 micro. In addition, BoW was used to generate text representations in the study. In [15], the study focused on multi-label prediction of ICD-9 codes from free-form EHR data. Using the MIMIC-III dataset with 18, 50, and 155 labels, several multi-label machine learning methods, including BR, CC, ECC, and MLkNN, were evaluated, with LR as the base classifier. FastText embeddings were used for text representation, and the model was retrained on the MIMIC-III and TREC 2017 medical corpora to examine the effect of domain-specific embeddings. Experimental results demonstrated that pre-trained high-dimensional FastText embeddings significantly improved performance, particularly for infrequent labels. Overall, the ECC–LR combination achieved the best results. In [13], an SVM model was used for the MLTC problem and enhanced with biomedical domain-specific features. In experiments conducted on the HoC dataset, biomedical entities such as DNA, RNA, cell types, and chemicals were extracted and these entities were converted into vector representations using BioWordVec. These representations were combined with MeSH embeddings to create document-level representations. The SVM model was developed using this representation structure is presented for the biomedical MLTC task.

2.2 Transformer-Based Approaches for MLTC

However, models based on transformer architectures, which have become popular in recent years, stand out for their potential to model deeper semantic relationships in multi-label medical texts through their contextual learning capabilities. Ref. [30] proposed a transformer-based deep learning model to assign multiple labels to a cancer article based on its content. The pre-trained BERT model was used to generate contextual embeddings. In their comparative analysis, the authors evaluated a "BERT + X", where X represents TextRNN, TextCNN, FastText, DPCNN, or DRNN. Among these, the BERT + TextRNN combination achieved the

best performance. In [31], the performance of encoder-only and encoder-decoder-based methods on multi-label classification was investigated. The study, conducted on four different open-source datasets, proposed the encoder-only methods Encoder+Head and LWAN, and the encoder-decoder methods Seq2Seq and T5Enc. Experimental results showed the superior performance of encoder-decoder methods, especially when used non-autoregressively. Ref. [12] presented an approach called QUAD-LLM-MLTC for multi-label classification. QUAD-LLM-MLTC includes the GPT-4o, BERT, PEGASUS, and BART models and leverages their strengths. Experimental results show that QUAD-LLM-MLTC yields a notable improvement in overall classification performance. Although general-purpose transformer models have achieved effective results on multi-label medical texts, given the domain-specific terminology of the biomedical domain, domain-specific models have enabled more accurate and contextually rich representations. In [26], pre-trained models were fine-tuned for multi-label text classification. Experimental results demonstrate improvements in the F1 score of domain-specific models BioMed-RoBERTa and PubMedBERT for fixed sequence length. Another study [25] examined multi-label classification to map clinicians' surgical notes to corresponding surgical procedure codes. TF-IDF and embedding-based approaches were applied to generate text representations. Using 350,000 surgical reports, the study examined the performance of MedBERT.de, SurgeryBERT.at, FastText, CNN, SVM, and LR models. MedBERT.de achieved the best performance, with a precision of 0.880. In [6], text representation methods such as TF-IDF, GloVe, Word2Vec, and fastText, as well as Transformer-based approaches including BERT, BioBERT, SciBERT, and PubMedBERT, are comparatively examined for multi-label classification. The study highlights PubMedBERT as the model with the best classification performance and further investigates the impact of text representation methods on overall classification effectiveness. In a study based on the BioBERT model [10], LitMC-BERT is proposed. The proposed approach learns label-specific representations for each label, as well as the co-occurrence relationships between label pairs. Experimental results on the LitCovid and HoC datasets demonstrate the effectiveness of the proposed method. Study [11] focused on addressing label imbalance, which is considered a significant challenge in multi-label classification. In the proposed MCICT approach, BioBERT is used to extract representations from texts, while the CoEAI-GCN module, presented as a Graph Convolutional Network-based module, both enriches label representations and constructs a correlation matrix by utilizing label co-occurrence relationships. Two real clinical datasets were used in the study, and experimental findings demonstrate improvements in classification performance. Ref. [32] proposed BioElectra-BiLSTM-Dual Attention classifier to address multi-label classification for biomedical literature. The proposed method employs the BioElectra encoder for feature extraction, utilizes BiLSTM to learn bidirectional contextual relationships, and applies a dual attention mechanism to emphasize important information. To improve MLTC performance, label embeddings and a label correlation matrix are used to provide additional information. Experiments conducted on the LitCovid dataset show that the model achieves a macro F1 score of 0.9165. In study [33], ensemble learning approach was used to adress MLTC problem. Model performance was improved by combining the outputs of different BERT-based models such as PubMedBERT, BioBERT, Bioformer and ELECTRA. The method was evaluated with experiments conducted on the LitCovid dataset. In [34], experimental studies were conducted on the HoC dataset, and a hybrid architecture was proposed for the MLTC task. In this study, contextual representations of the texts were extracted using the PubMedBERT model. And then these representations were combined with CNN layers.

The reviewed studies indicate that transformer-based pre-trained language models have substantially improved the performance of biomedical multi-label text classification through their contextual representations. However, most existing studies have focused on contextual representation learning, label dependency modeling, or architectural improvements, while comparatively less attention has been given to the joint use of topic modeling–based thematic information with these contextual representations.

2.3 Topic Modeling for Biomedical Text

Medical documents, with their terminological density, conceptual diversity, and structures involving multiple semantic contexts, provide a suitable data source for the use of topic modeling techniques. In [35], multi-label classification based on ICD codes was performed, and topic modeling representations were investigated in this context. In the study, LDA and its supervised variant, Partially Labelled LDA (PLDA), were applied to extract topic probability distributions as document representations, and the relationship between these representations and ICD labels was analyzed. The results showed that PLDA produced more effective topic models and achieved better performance when combined with classifiers. In [36], the study aimed to predict ICD-10 codes from cardiology EHR records in a multi-label classification. For this purpose, the multinomial distributions of topics obtained via LDA were used as document representations and served as inputs to supervised classifiers. While LDA-based methods play a significant role in the topic modeling literature, their context-independent representations limit their ability to capture semantic diversity. This has encouraged the inclusion of semantic representations from context-aware language models in topic modeling processes. Embedding-based topic modeling approaches enable modeling richer semantic relationships at the document level. For example, in the study [23], the results contributed to understanding the trends and changes in malaria research. In [37], a dual-perspective approach is presented to enhance cancer data text classification. In this approach, thematic clusters generated by BERTopic are compared with SVM classifier decisions, and BERTopic is evaluated as a component for assessing the consistency of classification outputs. In [38], it is emphasized that topics obtained through topic modeling can support the classification of electronic health record texts by enhancing interpretability and predictive performance. In this study, dimension reduction-based methods and Dirichlet-based probabilistic topic models are evaluated. The findings show that ProdLDA performs particularly well with respect to classification performance. In another study that compares topic modeling methods, the impact of various topic modeling approaches on text classification performance is evaluated within the TextNetTopics framework [39]. The study demonstrates that identifying the most suitable topic selection strategy is critical for improving text classification performance. Moreover, the proposed ENTM-TS (Ensemble Topic Modeling for Topic Selection) approach aims to generate more discriminative topic representations by grouping, scoring, and modeling topics obtained from multiple topic modeling methods, and the experimental results confirm the effectiveness of this approach.

These studies demonstrate that topic modeling is effective in uncovering latent thematic structures in biomedical texts and supporting a variety of natural language processing tasks. Nevertheless, studies that combine topic-derived thematic information with the contextual representations of pre-trained language models for biomedical multi-label text classification remain relatively limited. In this regard, jointly exploiting these two complementary types of representations represents a promising direction for biomedical MLTC.

The existing literature illustrates how text representation and modeling approaches in the biomedical MLTC task have evolved over time, and the main methods and their main contributions are summarized in Table 1.

Table 1: Overview of multi-label text classification studies in the biomedical domain.

Category	Approach/Model	Dataset	Main Contribution	Ref.
	BoW, MNB, KNN, SVM, LR, ECCLR, ECCSVM	Clinical documents	Ensemble of classifier chains achieved strong performance using BoW representations.	[29]

(Continued)

Table 1 (continued)

Category	Approach/Model	Dataset	Main Contribution	Ref.
Traditional Approaches for MLTC	FastText, BR, CC, ECC, MLkNN	MIMIC-III, TREC 2017	Domain-specific FastText embeddings improved prediction of infrequent ICD labels.	[15]
	SVM BioWordVec	HoC	SVM model was enhanced with biomedical domain-specific features for the MLTC problem.	[13]
Transformer-Based Approaches for MLTC	BERT + TextRNN	Cancer literature	Contextual embeddings significantly improved multi-label classification.	[30]
	Encoder-Head, LWAN, Seq2Seq, T5Enc	UKLEX, EURLEX, BIOASQ, MIMIC-III	Encoder–decoder architectures outperformed encoder-only models.	[31]
	QUAD-LLM-MLTC	HoC	Leveraging multiple LLMs improved overall classification accuracy.	[12]
	TextCNN, CAML, DR-CAML, BiGRU Longformer, BERT-base, ClinicalBERT, BioMed-RoBERTa, PubMedBERT	MIMIC-III, eICU	Analysis of domain-specific transformer based models and traditional neural networks for different sequence lengths.	[26]
	TF-IDF, FastText, MedBERT.de, SurgeryBERT.at, CNN, SVM, LR	Clinical Dataset	Multi-label classification was integrated into clinical workflows and demonstrated that SVM and FastText are hardware-efficient alternatives with high accuracy as BERT-based models.	[25]
	BERT, BioBERT, SciBERT, PubMedBERT	CEI, HoC	PubMedBERT achieved the best multi-label performance.	[6]

(Continued)

Table 1 (continued)

Category	Approach/Model	Dataset	Main Contribution	Ref.
	LitMC-BERT	LitCovid, HoC	Label-specific representations improved multi-label prediction.	[10]
	MCICT, CoEAI-GCN BioBERT, Graph convolutional network	Clinical Dataset	To obtain text representations and address label imbalance, BioBERT and graph convolutional network based method were used.	[11]
	BioElectra-BiLSTM-Dual Attention classifier	LitCovid	Bio-Electra encoder for feature extraction, utilizes BiLSTM to learn bidirectional contextual relationships, and applies a dual attention mechanism to emphasize important information.	[32]
	PubMedBERT, BioBERT Bioformer, ELECTRA	LitCovid	Ensemble methods was used to combine different PLMs.	[33]
	PubMedBERT CNN	HoC	Contextual representations were obtained using PubMedBERT and combined with CNN layers.	[34]
Topic Modeling for Biomedical Text	LDA, PLDA	ICD-coded EHR	PLDA produced more effective topic representations for classification.	[35]
	LDA	Cardiology EHR	Topic distributions used as features for supervised classifiers.	[36]
	BERTopic + SVM	Cancer texts	Topic clusters used to assess classification consistency.	[37]
	ProdLDA, CTM	EHR	ProdLDA showed strong predictive performance.	[38]
	TextNetTopics, ENTM-TS	WOS, CAMDA	Ensemble topic selection improved discriminative power.	[39]

(Continued)

Table 1 (continued)

Category	Approach/Model	Dataset	Main Contribution	Ref.
	CombinedTM + PLMs	**HoC**	**Integrates CombinedTM derived topic probability distributions with contextual PLM representations, enabling general-purpose PLMs to achieve competitive performance with domain-specific biomedical models while further enhancing biomedical PLMs.**	**This study**

3 Proposed Method

This study proposes an integrated modeling approach for multi-label classification of biomedical texts that combines contextual language representations of the text with topic information reflecting the high-level conceptual structure. Biomedical texts, particularly the cancer literature that is the focus of this study, have a multilayered, complex structure. Such text simultaneously contains information on multiple biological processes, molecular mechanisms, cellular functions, and treatment. This complex structure makes it challenging to base classification decisions solely on contextual representations. Therefore, the combined use of representations at different levels of abstraction, along with contextual representations, becomes essential. To address this requirement, the proposed approach is designed around two complementary components;

(i) A context-aware topic modeling approach is employed to extract topic probability distributions for each document. It aims to reveal the latent conceptual structures and dominant biological research orientations embedded in the text.

(ii) The topic probability distributions obtained from topic modeling are utilized as an additional representation layer for each document. This representation complements the contextual representations generated by pre-trained Transformer-based language models and is jointly used during the classification process.

Through this proposed structure, the transformer-based model is able to make classification decisions not only based on the contextual representation of the text but also by considering the conceptual position of the document in the context and its scientific orientations.

3.1 Contextualized Topic Model

To capture the latent topic of biomedical texts, a context-aware topic modeling approach is employed in this study. Specifically, the CombinedTM, a variant of the Contextualized Topic Model (CTM) framework, is adopted due to its ability to jointly leverage statistical word distributions and contextualized language representations [28]. CombinedTM can overcome the limited inference capacity of traditional topic modeling approaches that rely solely on statistical co-occurrence patterns by integrating contextualized language representations into the topic modeling process. In this way, words are modeled not only on their statistical co-occurrence frequencies but also on their contextual usage within the text. Considering the complex nature

of biomedical texts, CombinedTM is used in this work to obtain topic distributions that more accurately reflect the underlying scientific orientations and conceptual density of the documents.

3.1.1 Text Representation for CombinedTM

CombinedTM represents each document at two different levels of representation. With its dual representational design, the topic modeling process can utilize both statistical information derived from word distributions and semantic information obtained from contextual representations. Firstly, documents are represented using the BoW method, which gives word-level statistical information. This method enables learning word distributions within the document. Secondly, the same document is represented using pre-trained transformer-based language models to obtain contextual embeddings. These contextual embeddings reflect the latent meanings that words acquire depending on their position within the text and their interactions with surrounding words.

3.1.2 Topic Distribution Extraction

CombinedTM performs modeling by combining information obtained from contextual and statistical representations. The modeling results in the generation of topic probability distributions representing the topic structure for each document. These topic probability distributions represent the position of each document within the topic space learned by the model using a quantitative vector representation. This vector consists of probability values representing the relative importance of each topic for the relevant document within the topic space defined by CombinedTM. Topic distribution vectors are prepared for all data subsets to serve as additional text representations in the classification models. Topic modeling and classification are performed in a two-stage manner: CombinedTM is trained to generate topic distributions in the first stage, and its outputs are used as fixed inputs in the second stage.

The document collection is defined as

$$\mathcal{D} = \{d_1, d_2, \ldots, d_i, \ldots d_N\}, \quad i = 1, 2, \ldots, N, \tag{1}$$

where $\mathcal{D}$ denotes a corpus consisting of N biomedical documents (specifically, document abstracts), and d_i denotes the i-th document represented as an ordered sequence of textual tokens. In the multi-label text classification task, the objective is to learn a mapping from each document d_i to a binary label vector $\mathbf{y}_i \in \{0,1\}^L$, where $\mathbf{y}_i$ represents the multi-label assignments for the document and L denotes the total number of predefined labels.

For each document d_i, a topic probability distribution vector is extracted to represent its latent topical structure as

$$\boldsymbol{\theta}_i = \begin{bmatrix} \theta_{i,1} \\ \theta_{i,2} \\ \vdots \\ \theta_{i,K} \end{bmatrix}, \qquad \theta_{i,k} \in [0,1], \quad k = 1, \ldots, K, \tag{2}$$

where K denotes the total number of topics specified in the model, and each component $\theta_{i,k}$ represents the strength of association between document d_i and the k-th topic.

3.2 Enhancing MLTC with Topic Probability Distributions

In the proposed MLTC approach, an enriched representation is constructed for each document by jointly utilizing the outputs of a PLM and topic modeling. A document d_i is transformed into a contextual

text representation using a PLM. The document-level representation obtained from the final hidden layer of the model, presented as the [CLS] vector, is expressed as

$$\mathbf{h}_i^{\mathrm{CLS}} = \mathrm{PLM}_{\mathrm{CLS}}(d_i), \quad \mathbf{h}_i^{\mathrm{CLS}} \in \mathbb{R}^H, \tag{3}$$

where $\mathbf{h}_i^{\mathrm{CLS}} \in \mathbb{R}^H$ presents the contextual document representation of the i-th document obtained from the hidden layer of the PLM, and H represents the hidden dimension of the employed PLM.

The topic probability distribution vector $\boldsymbol{\theta}_i$, obtained from the CombinedTM model, is not directly fed into the classification model. Instead, it is processed through a learnable linear transformation followed by a non-linear activation function to obtain

$$\mathbf{h}_i^{\mathrm{topic}} = \mathrm{ReLU}\left(\mathbf{W}_t \boldsymbol{\theta}_i + \mathbf{b}_t\right), \quad \mathbf{h}_i^{\mathrm{topic}} \in \mathbb{R}^{128}, \tag{4}$$

where $\mathbf{W}_t \in \mathbb{R}^{128\times K}$ and $\mathbf{b}_t \in \mathbb{R}^{128}$ are the learnable weight matrix and bias vector of the topic transformation layer, respectively. The ReLU activation function offers nonlinearity and aims to enhance the discriminative power of the transformed topic representation. The resulting vector $\mathbf{h}_i^{\mathrm{topic}} \in \mathbb{R}^{128}$ corresponds to the latent topic representation of the document used as an additional input for the classification model. To minimize model complexity and maintain representation balance, the vector $\boldsymbol{\theta}_i$ has been transformed into a 128-dimensional space. The 128-dimensional hidden space was chosen to minimize the parameter load of the model while preserving the conceptual richness of the $\boldsymbol{\theta}_i$ vector. This allows the topic probability distribution to be integrated as an additional representation into the classification process without dominating contextual language representations.

The resulting contextual text representation and transformed topic representation are concatenated at the feature level to form a unified document representation:

$$\mathbf{h}_i^{\mathrm{enhanced}} = \left[\mathbf{h}_i^{\mathrm{CLS}}; \mathbf{h}_i^{\mathrm{topic}}\right] \in \mathbb{R}^{H+128}, \tag{5}$$

where $[\cdot;\cdot]$ denotes feature-level concatenation.

The enhanced representation $\mathbf{h}_i^{\mathrm{enhanced}}$ is passed through a dropout layer ($p = 0.2$) to improve the generalization capacity of the model and mitigate overfitting. The dropout rate was selected as 0.2 through hyperparameter optimization and kept fixed throughout all experiments to maintain a consistent training configuration. The resulting regularized representation is subsequently fed into the multi-label classification layer to compute the output logits as

$$\mathbf{z}_i = \mathbf{W}_{\mathrm{c}}\ \mathrm{Dropout}\left(\mathbf{h}_i^{\mathrm{enhanced}}\right) + \mathbf{b}_{\mathrm{c}}, \quad \mathbf{z}_i \in \mathbb{R}^L, \tag{6}$$

where $\mathbf{W}_{\mathrm{c}}$ and $\mathbf{b}_{\mathrm{c}}$ represent the learnable weight matrix and bias vector of the classification layer, respectively. This layer produces a logit vector $\mathbf{z}_i \in \mathbb{R}^L$, where L is the total number of labels. The j-th element of $\mathbf{z}_i$, denoted by $z_{i,j}$, represents the predicted logit associated with the j-th label of the i-th document.

In the MLTC task, Binary Cross-Entropy with logits loss (BCEWithLogitsLoss), which combines sigmoid activation and binary cross-entropy, was employed to generate independent binary decisions for each label. However, significant imbalances in label frequencies in datasets lead the model to focus on the majority classes and fail to adequately learn rare labels. To mitigate this problem, the standard BCEWithLogitsLoss in this study was improved with a label-frequency-based weighting mechanism. Following the weighting strategy suggested in [40], the positive class weight for the j-th label is computed, excluding the test data, as

$$pos\_weight_j = 1 - \frac{\sum_{i=1}^{N} y_{i,j}}{\sum_{i=1}^{N}\sum_{k=1}^{L} y_{i,k}}, \tag{7}$$

where $y_{i,j}$ denotes the ground truth label for the j-th class of the i-th document. The term $\sum_{i=1}^{N} y_{i,j}$ represents the total number of documents annotated with the label j, while $\sum_{i=1}^{N}\sum_{k=1}^{L} y_{i,k}$ represents the total number of positive label assignments on all documents and labels. By defining the positive weight as the complement of this ratio, the model assigns larger weights to infrequent labels and smaller weights to more common ones. This weighting approach helps the model focus more effectively on less common labels by limiting the tendency towards dominant labels during the learning process. The weighted BCEWithLogitsLoss is then formulated as

$$\mathcal{L}_i = -\frac{1}{L}\sum_{j=1}^{L}\Big[pos\_weight_j\, y_{i,j}\log(\sigma(z_{i,j})) + (1-y_{i,j})\log(1-\sigma(z_{i,j}))\Big]. \tag{8}$$

4 Experiments

This section presents the experimental setup of the proposed approach. Fig. 1 provides an overview of the experimental workflow, including data preparation, topic modeling, PLM-based multi-label classification, and evaluation.

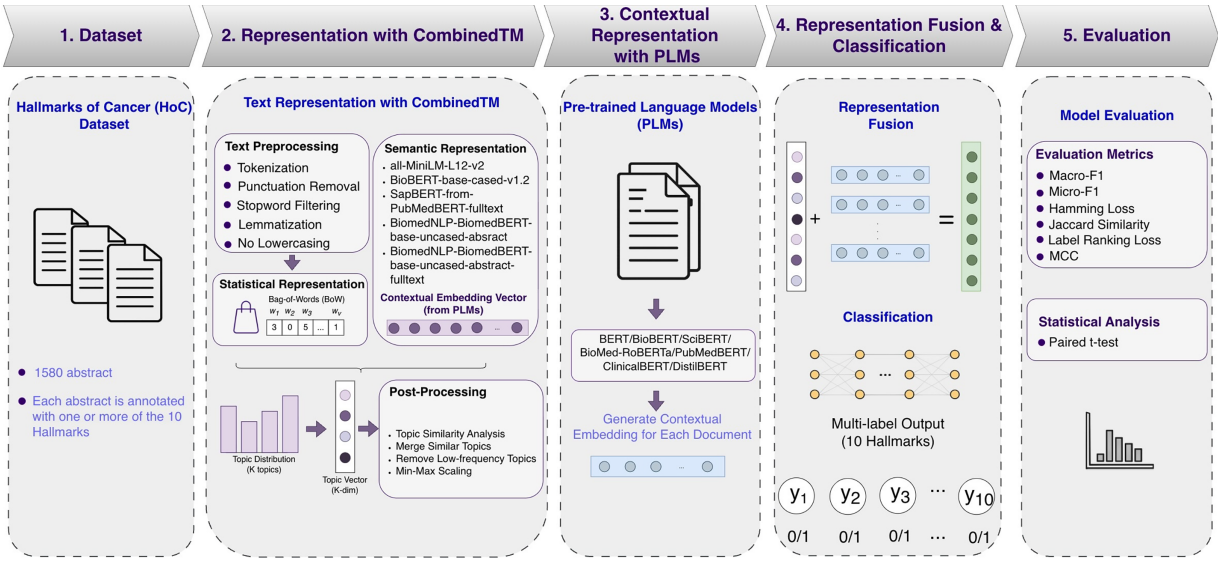

Figure 1: Illustration of the proposed topic-enhanced biomedical MLTC process.

4.1 Dataset

4.1.1 Hallmarks of Cancer (HoC)

The HoC dataset, proposed by Baker et al. [41], was developed to contribute to research on the automatic classification of cancer literature. The dataset consists of PubMed abstracts compiled to determine their relationship to known hallmarks of cancer. In this multi-label dataset, 10 hallmarks are defined for the abstracts, and each abstract is annotated with one or more hallmark categories. These hallmarks, introduced by Hanahan & Weinberg, are: *sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, genomic instability and mutation, tumor-promoting inflammation, cellular energetics, and avoiding immune destruction.*

The HoC dataset used in this study is the publicly available version cited in [42]. The train, validation, and test splits are provided to researchers in that study. This ensures the comparability of the results obtained during the modeling process with the findings in the literature. The train, validation, and test splits used in

this study comprise a total of 1580 abstracts, with an average of 1.56 labels per abstract. Among these abstracts, 919 are associated with one label, 478 with two labels, 145 with three labels, 31 with four labels, and 7 with five labels. Details about the dataset are provided in Fig. 2 and Table 2.

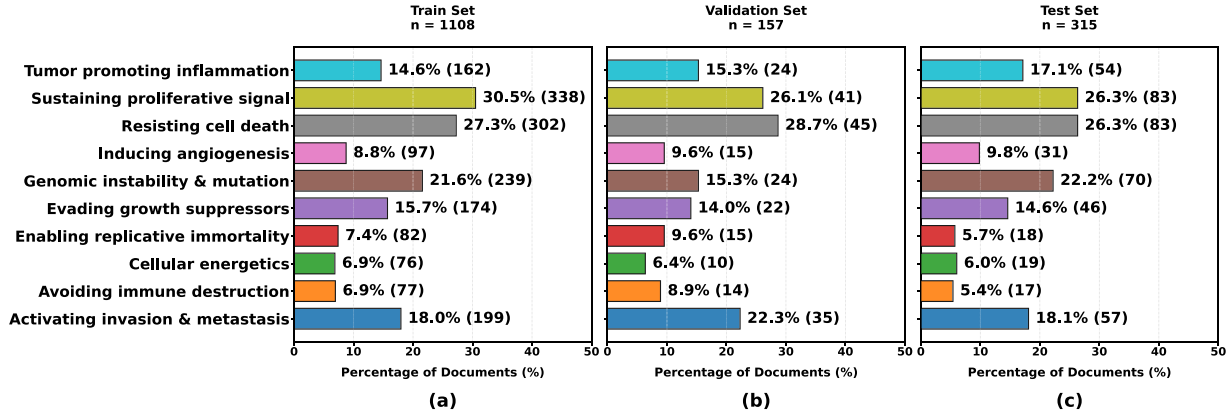

Figure 2: Hallmark distribution per dataset split: (**a**) train set, (**b**) validation set, and (**c**) test set.

Table 2: Statistics of the hallmarks of cancer dataset.

Dataset	Train/Val/Test	Total Abstracts	Unique Labels	Avg. Labels Per Abstract	Label Distribution (1/2/3/4/5)
HoC	1108/157/315	1580	10	1.56	919/478/145/31/7

4.1.2 Data Pre-Processing

Label Encoding

In the version of the HoC dataset used in this study [42], the raw labels are presented as strings separated by semicolons. Labels were converted into a binary matrix to make this structure suitable for model training. In this process, for each sample, the raw label string was parsed based on 10 predefined basic hallmarks of cancer. Then, separate columns were created for 10 unique classes in the dataset. Finally, a multi-label representation was obtained by assigning a value of '1' if the relevant class existed for each document, and '0' otherwise. As a result of this process, each document is represented by a binary vector of dimension $L = 10$, $\mathbf{y}_i \in \{0,1\}^L$.

Data Cleaning

Biomedical texts have a complex structure, characterized by dense use of abbreviations and punctuation, and meanings that may vary depending on case sensitivity. Following similar observations in the literature, the pre-processing steps in this study were kept to a minimum [43]. Given the PLMs' capabilities in the topic modeling and classification stages to handle such complex texts, only potential noise was removed. In this regard, newline characters and unnecessary multiple spaces, if present, were cleaned.

On the other hand, to reduce noise in the BoW-based statistical representation used in the CombinedTM framework, additional, more stringent preprocessing steps were applied. (i) First, texts are tokenized based on whitespace. This prevents the fragmentation and semantic distortion of special characters, parenthetical structures, and compound scientific expressions found in biomedical texts. (ii) To preserve

intra-term symbols, punctuation marks were not completely eliminated from the text, only punctuation characters appearing at the beginning or end of words (.,;:!? ’ ” “”), as well as isolated operators and parentheses not associated with any term were removed. In contrast, complex structures containing numbers, mathematical operators, or parentheses were treated as integral scientific units and preserved together with all their internal symbols. (iii) Subsequently, in addition to the standard stopword list, an additional stopword list was constructed for words frequently occurring in scientific articles, but non-discriminative terms such as “introduction,” “methods,” and “results,” and these were removed from the texts. As an exception, the token “T” was excluded from the stopword list to preserve meaningful biomedical expressions such as ‘T cell’. Other single-letter biological designations relevant to the dataset (e.g., “B” in “B cell”) were also examined, and it was confirmed that they are not included in the default NLTK English stopword list. (iv) To preserve terminology, tokens containing numbers, special characters, or symbols, as well as short tokens written entirely in uppercase, were considered scientific terms and retained without modification. For the remaining tokens, POS (Part of Speech)-based lemmatization was applied to reduce morphological variations. (v) In addition, lowercase conversion was not applied in the preprocessing step or in the BoW process, preserving case sensitivity.

These preprocessing steps minimized noise in the text for the statistical component of the topic modeling approach, while simultaneously preserving the ability of pretrained language models to directly learn natural language use.

4.2 Contextual Text Representation for Topic Modeling

In the contextual text representation part of the CombinedTM, both general-purpose and biomedical language models for the biomedical domain were employed. In this way, the impact of topic distributions derived from different embedding model-based representations on multi-label classification performance was experimentally demonstrated. Table 3 provides an overview of the embedding models used in the study.

Table 3: Embedding-oriented comparison of general-purpose and biomedical language models used in this study.

Model	Embedding Type	Embedding Size	Training Objective	Embedding Characteristics
all-MiniLM-L12-v2	Sentence-level dense embedding	384 (sentence)	Contrastive learning on large-scale multi-domain corpora	Efficient sentence embeddings, strong capability in capturing global semantic similarity; well-suited for contextual components in CTM.
BioBERT-base-cased-v1.2	Contextual token embedding	768 (token/pooled sentence)	Masked Language Modeling (MLM) on PubMed abstracts and PMC full-text	Domain-specific contextual embeddings with strong biomedical semantic awareness capture terminology-sensitive representations useful for biomedical tasks.

(Continued)

Table 3 (continued)

Model	Embedding Type	Embedding Size	Training Objective	Embedding Characteristics
SapBERT-from-PubMedBERT-fulltext	Concept-aligned contextual token embedding	768 (token/pooled sentence)	MLM + concept alignment (UMLS (Unified Medical Language System)-based)	Embedding space is semantically aligned across biomedical entities, reduces synonymy, and produces coherent and stable thematic clusters in topic modeling.
BiomedNLP-BiomedBERT-base-uncased-abstract	Contextual token embedding (abstract-level)	768 (token/pooled sentence)	MLM on PubMed abstracts	Enhanced in abstract-level biomedical semantics, embeddings reflect scientific terms commonly found in biomedical literature.
BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext	Contextual token embedding (abstract + full-text)	768 (token/pooled sentence)	MLM on PubMed abstracts and full-text articles	Rich contextual embeddings incorporating full text, include experimental findings and technical details, enabling comprehensive biomedical knowledge representation.

all-MiniLM-L12-v2: As a general-purpose language model within the Sentence-Transformers, all-MiniLM-L12-v2 is based on the MiniLM architecture [44]. This model effectively captures semantic similarities thanks to its training on large-scale, multi-domain datasets.

BioBERT-base-cased-v1.2: Built on the BERT architecture, BioBERT-base-cased-v1.2 is a language model fully adapted to the biomedical domain, pre-trained on a large-scale biomedical corpus [45].

SapBERT-from-PubMedBERT-fulltext: Built upon the PubMedBERT architecture, this model is enhanced through additional training based on concept alignment tasks designed to strengthen semantic similarity among biomedical named entities [46]. It aims to reduce synonymy and conceptual diversity in biomedical terminology, ensuring that different nomenclatures for the same medical concept are positioned closer together in the semantic space. In this respect, it is a model used to obtain more consistent thematic clusters in topic modeling processes.

BiomedNLP-BiomedBERT-base-uncased-abstract/fulltext: These models are introduced as domain-specific models pre-trained on PubMed abstracts and both PubMed abstracts and full-text articles, respectively [47]. Since the training corpus in the abstract version consists solely of scientific abstracts, it demonstrates improved alignment with terminology commonly used in biomedical research literature

compared to general-purpose models. On the other hand, the full-text version is trained not only on PubMed abstracts but also on training full-text. It is therefore enriched with methodological descriptions, experimental findings, discussions, and technical details, enabling a more comprehensive representation of biomedical knowledge.

4.3 Topic Post-Processing and Refinement

After topic modeling, a K-dimensional document-topic probability distribution is obtained for each document. The resulting raw topic distributions may include topics that are highly similar or are used very rarely in the dataset. This can reduce the interpretability of the model and introduce noise into the representations used in subsequent stages. To prevent this, a check is performed on the obtained topic distributions through post-processing.

Using the resulting document-topic matrix, each topic was represented by a vector that reflects its distribution across all documents in the dataset. By computing cosine similarity between these vectors, topics with a highly similar distribution were identified. Topic pairs with a similarity value above the threshold (0.90) were considered to be semantically redundant. This threshold was determined through experimental evaluation to ensure that only topics exhibiting a very high degree of similarity were merged, thereby minimizing the risk of combining semantically distinct topics. In such cases, the topic with the lower index was selected as the representative, and similar topics were grouped under it. Subsequently, to assess how extensively each topic is used across the dataset, topic usage rates are computed by taking the average of the document–topic probabilities over the training set. Topics whose average usage falls below a predefined threshold (0.01) are considered low-frequency topics that are insufficiently represented in the dataset; therefore, they are excluded from further analysis. Similarly, this threshold was determined through experimental evaluation to ensure that only topics with negligible representation in the dataset were removed. As a result of these two steps, redundant topics are merged, and low-usage topics are removed, yielding a more compact set of topics. During the merging process, probability values of topics grouped under the same representative were summed to create new document-topic distributions, and the sum of probabilities for each document was renormalized to 1 again. Finally, the resulting new document-topic distributions were scaled to the range $[0, 1]$ using Min-Max normalization to ensure consistent scaling across different data segments. This process enabled the topic representations to be used more consistently and comparably in subsequent classification and analysis stages.

4.4 Pre-Trained Language Models for MLTC

The adaptation of PLMs to biomedical datasets for domain-specific purposes has improved their performance in biomedical research and further expanded their application domains. This subsection will provide a detailed overview of the pre-trained general-purpose and domain-specific models evaluated in this study (Table 4).

4.4.1 General-Purpose Models

This group includes language models pretrained on general-domain corpora without domain-specific adaptation.

Table 4: Summary of pre-trained language models.

Model	Training Data	Vocabulary	Key Features
google-bert/bert-base-cased	Wikipedia, BookCorpus	WordPiece	First bidirectional Transformer PLM.
distilbert/distilbert-base-cased	Wikipedia, BookCorpus	WordPiece	Lightweight BERT distilled via knowledge distillation.
dmis-lab/biobert-v1.1	PubMed abstracts, PMC full texts	WordPiece	Biomedical domain adaptation.
medicalai/ClinicalBERT	Clinical records	WordPiece	Adapted to clinical jargon and abbreviations.
microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract (PubMedBERT)	PubMed abstracts	WordPiece + Whole Word Masking	From-scratch biomedical pretraining.
allenai/scibert_scivocab_cased	1.14M scientific papers (82% biomed, 18% CS)	SciVocab	Scientific-domain vocabulary.
allenai/biomed_roberta_base	Biomedical corpora	RoBERTa BPE	RoBERTa backbone with dynamic masking.

BERT: BERT (Bidirectional Encoder Representations from Transformers) model, proposed by Devlin et al. stands out with its bidirectional contextualization based on the Transformer architecture [48]. Through two pre-training tasks, Masked Language Modeling and Next Sentence Prediction, the model learns representations by considering both the left and right contexts of words. With this structure, the model generates context-dependent word representations. Its success across tasks such as classification, question-answer, and named entity recognition has led to the development of new domain-specific variants of this model.

DistilBERT: DistilBERT is a Transformer-based language model developed using knowledge distillation to provide the contextual representation capability of BERT at a significantly lower computational cost [49]. Trained within a teacher–student learning framework, the model aims to capture BERT's learned linguistic knowledge using fewer layers and parameters. This significantly reduces model size and inference time while maintaining competitive performance in many natural language processing tasks.

4.4.2 Domain-Adaptive Models

This group consists of language models pretrained or adapted on biomedical-domain corpora.

BioBERT: BioBERT, proposed by Lee et al. [45], is based on the general-purpose BERT model. To present a language model adapted to biomedical texts, additional pre-training was performed on large-scale biomedical texts, including PubMed abstracts and full-text articles from PubMed Central. This model represents a significant advance in the field, as it is one of the first large-scale pre-trained language models explicitly developed for biomedical research.

ClinicalBERT: ClinicalBERT is a language model adapted to clinical texts [50]. Trained using large-scale clinical records, the model aims to better capture clinical terminology, contextual relationships, and

linguistic characteristics found in patient records. Its adaptability to the complex and abbreviated nature of medical language makes it a popular model for biomedical research. Studies in the literature demonstrate that ClinicalBERT outperforms general-purpose BERT in tasks such as clinical note analysis, patient outcome prediction, diagnostic code classification, and medical decision support systems.

PubMedBERT: PubMedBERT (absract) was presented as a model based on the BERT architecture, trained from scratch with biomedical texts obtained from PubMed articles [47]. Unlike general-purpose BERT models, or models that are pre-trained on general text and then adapted with biomedical data, it is trained entirely on data derived from biomedical literature, providing a deeper alignment with biomedical terminology. In training, a massive biomedical dataset consisting of PubMed abstracts was used.

SciBERT: SciBERT, proposed by Beltagy et al., is another model based on the BERT architecture and is a language model pre-trained on scientific articles [51]. The model was trained using 1.14 million full-text scientific articles, 82% of which are from biomedical science and 18% from computer science. In addition to having a specialized training corpus, SciBERT also uses a domain-specific vocabulary, SciVocab, enabling it to better represent frequently used terms and domain-specific concepts in the scientific literature. Comparative studies have shown that SciBERT outperforms general-purpose BERT on tasks based on scientific texts.

BioMed-RoBERTa: RoBERTa is based on the BERT architecture, an improved model trained on larger datasets, for longer training durations, and dynamic masking strategies. BioMed-RoBERTa is a domain-specific model built on RoBERTa and pre-trained on biomedical data [52]. BioMed-RoBERTa was trained on extensive text from biomedical corpora. This specialization can more accurately represent biomedical terms, their contexts, and their use in the literature.

4.5 Evaluation Metrics

In the literature, various metrics such as F1-score, MCC, Hamming Loss, Jaccard Similarity, and Ranking Loss are used to evaluate the performance of multi-label classification.

4.5.1 F1-Score

F1-score is a metric that measures the balance between precision and recall. Precision and recall represent the proportion of the model's positive predictions that are actually correct and the proportion of actual positive examples that the model correctly identifies, respectively. The F1-score is the harmonic mean of precision and recall and is widely used to evaluate classification performance, particularly for datasets with imbalanced class distributions. It is defined as follows:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{9}$$

F1-score variants, Macro-F1, Micro-F1, and Weighted-F1, are widely used metrics to evaluate classification performance. Macro-F1 calculates the F1-score for each class and takes the arithmetic mean of the scores. In cases of data imbalance, it reduces the impact of majority classes and provides a better reflection of the performance on minority classes. Micro-F1 is calculated based on the total true positives, false positives, and false negatives across all classes. It is more sensitive to the performance of the majority classes with a large number of samples. Weighted-F1, on the other hand, calculates the F1-score for each class and then takes the mean of the classes weighted by the number of samples. This gives greater weight to classes with a larger number of samples in the presence of data imbalance.

4.5.2 Matthews Correlation Coefficient (MCC)

MCC is an evaluation metric used to assess the performance of classification models by jointly considering true positives, true negatives, false positives, and false negatives [53]. By considering all prediction outcomes, MCC provides a reliable measure of classification performance, particularly for datasets with imbalanced class distributions. Therefore, it is widely used for evaluating classification performance in problems involving imbalanced data distributions [54].

4.5.3 Hamming Loss

Hamming Loss is a metric used to evaluate multi-label classification performance, measuring the average difference between the predicted and true label sets. For each sample, the predicted labels are compared with the true labels, and the proportion of incorrectly predicted labels is calculated. The measured value ranges from 0 to 1, with 0 indicating that all labels were correctly predicted. It is defined as follows:

$$\text{Hamming-Loss} = \frac{1}{N \cdot L} \sum_{i=1}^{N} \sum_{j=1}^{L} I\left(y_{ij} \neq \hat{y}_{ij}\right). \tag{10}$$

In Eq. (10), N is total number of instances, L is number of labels, y_{ij} presents the true value of the j-th label for the i-th instance, $\hat{y}_{ij}$ presents predicted value of the j-th label for the i-th instance and $I(\cdot)$ is indicator function, which returns 1 if the condition holds, and 0 otherwise.

4.5.4 Jaccard Similarity

Jaccard Similarity is a metric used to measure the similarity between two clusters. In multi-label classification, this metric measures performance based on the ratio of the intersection to the union of the predicted and true label sets for an instance. Its value ranges from 0 to 1, where 1 indicates complete overlap between the predicted and true label sets. It is defined as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{11}$$

where A represents the true label set and B represents the label set predicted by the model.

Accordingly, the experimental results are reported using Jaccard Micro and Jaccard Macro.

4.5.5 Ranking Loss

Label Ranking Loss measures the extent to which the true labels for an instance are incorrectly positioned in the predicted score ranking. Its value ranges from 0 to 1, where 0 indicates that all relevant labels are ranked above the irrelevant labels. It is defined as follows:

$$\text{Ranking-Loss} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|Y_i| \cdot |\bar{Y}_i|} \times \left|\left\{(y^+, y^-) : f_i(y^+) \leq f_i(y^-)\right\}\right|, \tag{12}$$

where N is the number of instances, L is the number of labels in each instance, Y_i presents the set of true positive labels for the i-th instance, $\bar{Y}_i$ presents the set of true negative labels for the i-th instance, and $f_i(y)$ is the score predicted by the model for label y of the i-th instance.

4.6 Implementation Details and Parameters

Table 5 summarizes all components and related parameters used in the topic modeling phase. The scikit-learn CountVectorizer was used to construct a statistical BoW representation of the text. Vectorization was performed using only the vocabulary learned from the training set; validation and test sets were transformed using this fixed vocabulary. The lowercase was set as false to maintain case sensitivity. For contextual representations, the Sentence Transformers (SBERT) library was used, and the models were obtained from Hugging Face. The models used, and their corresponding embedding dimensions, are detailed in Table 3. In the topic modeling performed with CombinedTM, the number of topics was configured as (K = 15) and trained over 50 epochs with a mini-batch size of 64. During the inference stage for obtaining document–topic probability distributions, the sample size was set to n = 20. The parameters used in the post-processing steps applied to the resulting document–topic distributions are summarized in Table 5.

Table 5: Configurations were used in the topic modeling stage.

Component	Parameter	Value
BoW Rep.	Vectorizer	CountVectorizer
	Input Text	Preprocessed Text
	Lowercase	False
	Vocab Source	Training Split Only
Contextual Rep.	Encoder Model	Embedding Models*
	Input Text	Raw Text
	Dimension	Embedding Size*
Topic Model	Model Type	CombinedTM
	Topics (K)	15
	Epochs	50
	Batch Size	64
Inference	Samples (n)	20
	Output	Topic Probability Distributions
Post-proc.	Similarity	Cosine Similarity
	Threshold	0.90
	Usage Limit	0.01
	Scaling	Min–Max Normalization

Note: *Refer to Table 3 for details on embedding models and embedding size.

Table 6 summarizes the model fine-tuning and optimization parameters used in the classification stage. During the fine-tuning process, the learning rate was set to 1e−4, and the models were trained for 3 epochs. A mini-batch size of 8 was used for both fine-tuning and evaluation. To reduce overfitting, the weight decay parameter was set to 0.01. The maximum input sequence length was limited to 512 tokens due to the architectural constraints of PLM. For the multi-label classification task, a sigmoid activation function was employed, and the classification threshold was set to 0.5. As the loss function, a weighted BCEWithLogitsLoss was adopted to address class imbalance. PLM performance was evaluated at the end of each epoch. PLMs used in this study were obtained from Hugging Face, and detailed information about these models is provided in Table 4.

Table 6: Model fine-tuning parameters.

Parameter	Value
Learning Rate	1e−4
Training Epochs	3
Batch Size (Train/Eval)	8/8
Weight Decay	0.01
Max Sequence Length	512
Classification Threshold	0.5 (Sigmoid)
Loss Function	Weighted BCEWithLogitsLoss
Evaluation Strategy	Epoch

4.7 Hardware Environment

All stages of the study, including data pre-processing, model fine-tuning, and evaluation, were conducted using the Google Colab environment. Since the pre-processing steps and non–computational–intensive components of the pipeline require relatively low computational resources, these parts were executed on the CPU. In contrast, the fine-tuning and evaluation of PLMs require greater computational capacity; therefore, these processes were performed using a T4 GPU.

5 Results and Discussion

In the proposed study, the MLTC problem is addressed using the HoC dataset, which has been specifically prepared for the cancer literature. The first stage aimed to fine-tuned general-purpose and biomedical domain-adaptive PLMs for specific tasks on this dataset. To ensure a fair comparison, all PLMs are fine-tuned under the same training (Fig. 2a), validation (Fig. 2b), and test (Fig. 2c) splits. In this way, the MLTC performance of general-purpose models and biomedical models developed with different pre-training strategies and training corpora is comparatively analyzed within the context of the HoC dataset. The Macro/Micro/Weighted-F1 Score, MCC, Hamming Loss, Jaccard Micro/Macro, and Label Ranking Loss metrics were used in the evaluations performed on the test dataset. To improve the reliability of the experimental results, all experiments were repeated over 10 runs, and the results were reported as mean values. The findings reveal significant performance differences between general-purpose and domain-adaptive PLMs on the HoC dataset. Despite their strong contextual representation capabilities, general-purpose models demonstrated limited performance when confronted with the dense biomedical terminology and the high degree of conceptual diversity inherent in the domain. In contrast, pre-trained domain-adaptive models on biomedical texts have been observed to produce more consistent results, particularly for metrics sensitive to label imbalance and ranking, such as Macro-F1 and Label Ranking Loss. This outcome can be considered expected given the imbalanced label distribution present in the HoC dataset (Fig. 2). In the MLTC scenario based solely on contextual representations, PubMedBERT (microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract) appears to be the most successful model. According to the results of baseline models presented in Table 7, PubMedBERT achieved Micro/Macro/Weighted-F1 scores of 0.8732/0.8755/0.8724, while also exhibiting the best performance in Hamming Loss of 0.0378, Jaccard Micro/Macro of 0.7751/0.7835, and Label Ranking Loss of 0.0147 metrics. These results indicate that PubMedBERT, a model trained from scratch on biomedical texts, better captures the conceptual structure and terminological relationships in the HoC dataset. Similarly, it is observed that the domain-adaptive models used in this study, owing to their domain-specific design, achieve more accurate and consistent performance than general-purpose models on HoC dataset, which is dense biomedical terminology. This finding suggests that adaptation to biomedical corpora,

along with domain-specific pre-training strategies and vocabularies, contributes more effective modeling of conceptual relationships.

During the fine-tuning stage, each PLM represents text in a context-dependent manner. However, a decision mechanism that relies solely on the contextual representations provided by the PLM may remain limited when explicit signals about the thematic orientation of the document are unavailable to the model. This limitation becomes more pronounced, particularly for general-purpose models, due to the diversity of biomedical terminology and the presence of conceptual synonymy. At this point, in the second stage of the study, topic modeling was incorporated as an additional representation method and combined with the contextual representations produced by the PLM to enhance classification performance. Considering that biomedical texts are characterized by dense terminology and high expression variability, and that topic modeling requires a modeling mechanism capable of overcoming the limitations of purely statistical approaches, the CombinedTM architecture was employed for topic modeling. CombinedTM preserves the statistical signal from word distributions in the BoW representation while simultaneously integrating dense contextual representations from embedding models into the topic modeling process via SBERT-based embedding extraction. An important point to emphasize is that, in this study, the SBERT library was employed as an architectural framework to obtain semantic representations of documents. This context-aware design of CombinedTM generates more meaningful topic distributions that account not only for frequently occurring lexical patterns but also for the context-dependent use of biomedical concepts. This study examines the contribution of topic modeling to the MLTC process within a comparative analysis framework. For this purpose, different embedding models presented in Table 3 were used in the contextual representation component of the CombinedTM architecture. Thus, while evaluating whether the obtained topic probability distributions truly provide additional representational power, it became clearer how the contributions of general-purpose and domain-adaptive language models to the contextual channel of CombinedTM are reflected in MLTC performance.

In this study, CombinedTM was trained over 50 epochs and initially configured to generate a 15-dimensional topic probability distribution for each document. However, directly using the raw outputs of the topic modeling process may, in some cases, result in topics that are either semantically overlapping or insufficiently represented across the dataset. To control this, post-processing steps detailed in Section 4.3 were applied. Post-processing analyses revealed that the effective number of topics does not remain fixed across different embedding configurations and runs; while it remains at the initially specified value of 15 in some cases, it falls below 15 in others. This variation stems from the stochastic nature of the topic modeling process and differences in how embedding methods separate the topic space. Despite these variations in the number of topics across runs under the same configuration, the fact that the classification performance remains within similar ranges demonstrates the robustness of the proposed approach against variations in the topic structure. Through the applied post-processing steps, ineffective or potentially noisy topic representations were eliminated, resulting in more semantically coherent and discriminative topic probability distributions. These refined topic distributions were integrated as additional representations into the contextual representations generated by the PLMs and used during the fine-tuning process for MLTC. Across multiple runs, the consistent contribution of topic probability distributions to classification performance indicates that the proposed approach is stable and reliable.

Table 7: Performance comparison of classifiers with different embedding models on the HoC dataset.

Classifiers	Embedding Models	F1-Score (Micro) mean/max/std	F1-Score (Macro) mean/max/std	F1-Score (Weighted) mean/max/std	MCC mean/max/std	Hamming Loss mean/min/std	Jaccard Similarity Micro mean/max/std	Jaccard Similarity Macro mean/max/std	Label Ranking Loss mean/min/std
BERT	Baseline Model	0.7870/0.8149/0.0291	0.7604/0.8072/0.0544	0.7713/0.8106/0.0378	0.7422/0.7806/0.0530	0.0587/0.0521/0.0068	0.6497/0.6876/0.0387	0.6405/0.6896/0.0551	0.0365/0.0277/0.0115
	all-MiniLM-L12-v2	0.8384/0.8620/0.0233	0.8405/0.8705/0.0223†	0.8351/0.8614/0.0250	0.8156/0.8464/0.0241	0.0468/0.0403/0.0061	0.7224/0.7575/0.0341	0.7324/0.7749/0.0297	0.0254/0.0177/0.0059
	BioBERT-base-cased-v1.2	0.8265/0.8565/0.0323	0.8231/0.8591/0.0394†	0.8209/0.8555/0.0390	0.8068/0.8339/0.0332	0.0496/0.0422/0.0073	0.7055/0.7491/0.0461	0.7105/0.7574/0.0484	0.0276/0.0203/0.0091
	SapBERT-from-PubMedBERT-fulltext	0.8345/0.8629/0.0207	0.8314/0.8615/0.0253†	0.8312/0.8614/0.0231	0.8037/0.8376/0.0284	0.0478/0.0410/0.0049	0.7165/0.7589/0.0305	0.7194/0.7590/0.0335	0.0290/0.0219/0.0063
	BiomedNLP-BiomedBERT-abstract	**0.8469/0.8624/0.0130**	**0.8458/0.8619/0.0163†**	**0.8445/0.8597/0.0145**	**0.8206/0.8371/0.0166**	**0.0450/0.0406/0.0033**	**0.7345/0.7580/0.0194**	**0.7397/0.7637/0.0226**	**0.0243/0.0196/0.0037**
	BiomedNLP-BiomedBERT-abstract-fulltext	0.8420/0.8654/0.0149	0.8405/0.8675/0.0205†	0.8393/0.8648/0.0170	0.8162/0.8439/0.0193	0.0460/0.0400/0.0037	0.7273/0.7627/0.0222	0.7313/0.7686/0.0253	0.0250/0.0156/0.0071
DistilBERT	Baseline Model	0.8164/0.8295/0.0110	0.8149/0.8345/0.0143	0.8095/0.8263/0.0136	0.7927/0.8121/0.0126	0.0512/0.0476/0.0025	0.6898/0.7087/0.0156	0.6996/0.7237/0.0179	0.0306/0.0261/0.0027
	all-MiniLM-L12-v2	**0.8401/0.8568/0.0142**	**0.8413/0.8620/0.0185†**	**0.8368/0.8553/0.0159**	**0.8163/0.8374/0.0199**	**0.0464/0.0419/0.0042**	**0.7245/0.7495/0.0211**	**0.7341/0.7629/0.0251**	**0.0253/0.0213/0.0039**
	BioBERT-base-cased-v1.2	0.8351/0.8628/0.0195	0.8347/0.8666/0.0204†	0.8312/0.8621/0.0207	0.8099/0.8447/0.0219	0.0474/0.0397/0.0049	0.7172/0.7587/0.0285	0.7250/0.7682/0.0281	0.0267/0.0212/0.0047
	SapBERT-from-PubMedBERT-fulltext	0.8259/0.8630/0.0220	0.8283/0.8642/0.0239	0.8224/0.8630/0.0228	0.8022/0.8407/0.0250	0.0499/0.0403/0.0055	0.7039/0.7590/0.0319	0.7160/0.7633/0.0338	0.0283/0.0199/0.0048
	BiomedNLP-BiomedBERT-abstract	0.8320/0.8651/0.0242	0.8303/0.8701/0.0321	0.8272/0.8647/0.0266	0.8063/0.8462/0.0308	0.0481/0.0400/0.0056	0.7130/0.7623/0.0349	0.7203/0.7741/0.0430	0.0290/0.0235/0.0042
	BiomedNLP-BiomedBERT-abstract-fulltext	0.8337/0.8541/0.0173	0.8333/0.8577/0.0199†	0.8305/0.8535/0.0188	0.8077/0.8323/0.0210	0.0480/0.0429/0.0045	0.7152/0.7453/0.0254	0.7218/0.7540/0.0268	0.0272/0.0203/0.0059
BioMed-RoBERTa	Baseline Model	0.8612/0.8786/0.0099	0.8665/0.8855/0.0112	0.8593/0.8775/0.0110	0.8408/0.8645/0.0118	0.0409/0.0359/0.0025	0.7564/0.7835/0.0152	0.7707/0.7972/0.0155	0.0170/0.0132/0.0026
	all-MiniLM-L12-v2	0.8645/0.8729/0.0079	0.8710/0.8830/0.0103	0.8633/0.8714/0.0087	0.8497/0.8605/0.0075	0.0404/0.0375/0.0019	0.7614/0.7744/0.0121	0.7771/0.7968/0.0160	0.0162/0.0142/0.0014
	BioBERT-base-cased-v1.2	0.8660/0.8774/0.0096	0.8692/0.8819/0.0088	0.8653/0.8772/0.0099	0.8460/0.8585/0.0096	0.0400/0.0368/0.0025	0.7637/0.7815/0.0147	0.7733/0.7935/0.0133	0.0169/0.0130/0.0023
	SapBERT-from-PubMedBERT-fulltext	0.8714/0.8810/0.0072	0.8779/0.8869/0.0074†	0.8710/0.8804/0.0074	0.8552/0.8654/0.0083	0.0387/0.0362/0.0020	0.7722/0.7873/0.0112	0.7867/0.8008/0.0120	0.0165/0.0136/0.0019
	BiomedNLP-BiomedBERT-abstract	**0.8755/0.8912/0.0073**	**0.8819/0.8945/0.0060†**	**0.8751/0.8913/0.0075**	**0.8546/0.8652/0.0091**	**0.0375/0.0330/0.0020**	**0.7786/0.8038/0.0116**	**0.7936/0.8128/0.0097**	**0.0152/0.0113/0.0018**
	BiomedNLP-BiomedBERT-abstract-fulltext	0.8686/0.8817/0.0070	0.8746/0.8905/0.0081	0.8680/0.8819/0.0073	0.8512/0.8689/0.0088	0.0395/0.0359/0.0018	0.7678/0.7884/0.0109	0.7826/0.8067/0.0127	0.0166/0.0132/0.0017
ClinicalBERT	Baseline Model	0.8124/0.8328/0.0119	0.8066/0.8297/0.0176	0.8025/0.8300/0.0186	0.7853/0.8043/0.0133	0.0523/0.0473/0.0028	0.6842/0.7135/0.0171	0.6922/0.7124/0.0146	0.0312/0.0235/0.0053
	all-MiniLM-L12-v2	0.8424/0.8565/0.0091	0.8461/0.8587/0.0084†	0.8390/0.8555/0.0102	0.8036/0.8367/0.0197	0.0456/0.0422/0.0021	0.7278/0.7491/0.0135	0.7409/0.7567/0.0108	0.0239/0.0199/0.0027
	BioBERT-base-cased-v1.2	0.8346/0.8636/0.0220	0.8357/0.8654/0.0298†	0.8298/0.8628/0.0275	0.7944/0.8292/0.0302	0.0477/0.0403/0.0054	0.7167/0.7599/0.0318	0.7283/0.7673/0.0365	0.0265/0.0196/0.0046
	SapBERT-from-PubMedBERT-fulltext	0.8454/0.8596/0.0126	0.8509/0.8702/0.0144†	0.8430/0.8589/0.0144	0.8108/0.8371/0.0175	0.0448/0.0413/0.0031	0.7324/0.7538/0.0188	0.7470/0.7746/0.0199	0.0235/0.0210/0.0025
	BiomedNLP-BiomedBERT-abstract	0.8321/0.8528/0.0153	0.8286/0.8613/0.0271†	0.8242/0.8514/0.0218	0.7851/0.8241/0.0284	0.0483/0.0432/0.0036	0.7127/0.7434/0.0224	0.7217/0.7597/0.0314	0.0281/0.0230/0.0035
	BiomedNLP-BiomedBERT-abstract-fulltext	**0.8464/0.8584/0.0112**	**0.8528/0.8615/0.0108†**	**0.8441/0.8573/0.0124**	**0.8157/0.8323/0.0131**	**0.0446/0.0413/0.0029**	**0.7339/0.7519/0.0166**	**0.7499/0.7615/0.0129**	**0.0234/0.0180/0.0029**

(Continued)

Table 7 (continued)

Classifiers	Embedding Models	F1-Score (Micro) mean/max/std	F1-Score (Macro) mean/max/std	F1-Score (Weighted) mean/max/std	MCC mean/max/std	Hamming Loss mean/min/std	Jaccard Similarity Micro mean/max/std	Jaccard Similarity Macro mean/max/std	Label Ranking Loss mean/min/std
PubMedBERT	Baseline Model	0.8732/0.8896/0.0131	0.8755/0.8915/0.0167	0.8724/0.8885/0.0133	0.8487/0.8711/0.0289	0.0378/0.0337/0.0032	0.7751/0.8011/0.0204	0.7835/0.8077/0.0260	0.0147/0.0109/0.0029
	all-MiniLM-L12-v2	0.8807/0.8912/0.0091	0.8846/0.8967/0.0112	0.8807/0.8912/0.0089	0.8627/0.8764/0.0128	0.0363/0.0330/0.0028	0.7871/0.8038/0.0144	0.7978/0.8158/0.0168	0.0127/0.0106/0.0014
	BioBERT-base-cased-v1.2	0.8823/0.8976/0.0071	0.8866/0.8999/0.0073	0.8821/0.8975/0.0071	0.8650/0.8808/0.0084	0.0357/0.0314/0.0021	0.7895/0.8143/0.0114	0.8007/0.8228/0.0119	0.0129/0.0107/0.0013
	SapBERT-from-PubMedBERT-fulltext	0.8836/0.8986/0.0089	0.8888/0.9012/0.0090†	0.8835/0.8986/0.0085	0.8674/0.8819/0.0106	0.0353/0.0308/0.0028	0.7916/0.8159/0.0143	0.8041/0.8258/0.0147	0.0135/0.0114/0.0016
	BiomedNLP-BiomedBERT-abstract	0.8815/0.8928/0.0067	0.8882/0.9048/0.0095†	0.8815/0.8924/0.0066	0.8664/0.8848/0.0106	0.0361/0.0327/0.0020	0.7883/0.8064/0.0108	0.8036/0.8309/0.0156	0.0132/0.0110/0.0013
	BiomedNLP-BiomedBERT-abstract-fulltext	**0.8839/0.8988/0.0084**	**0.8900/0.9026/0.0079†**	**0.8838/0.8991/0.0083**	**0.8685/0.8835/0.0094**	**0.0353/0.0311/0.0026**	**0.7921/0.8161/0.0134**	**0.8063/0.8262/0.0127**	**0.0130/0.0106/0.0013**
BioBERT-v1.1	Baseline Model	0.8682/0.8844/0.0085	0.8703/0.8896/0.0112	0.8670/0.8826/0.0092	0.8478/0.8697/0.0116	0.0391/0.0343/0.0023	0.7671/0.7927/0.0132	0.7750/0.8050/0.0166	0.0170/0.0147/0.0018
	all-MiniLM-L12-v2	0.8771/0.8907/0.0094	0.8852/0.9024/0.0091†	0.8769/0.8907/0.0097	0.8631/0.8815/0.0099	0.0369/0.0333/0.0024	0.7812/0.8030/0.0149	0.7984/0.8260/0.0144	0.0159/0.0138/0.0014
	BioBERT-base-cased-v1.2	0.8747/0.8877/0.0094	0.8783/0.8936/0.0113	0.8744/0.8881/0.0096	0.8560/0.8725/0.0124	0.0376/0.0343/0.0026	0.7775/0.7981/0.0147	0.7877/0.8111/0.0172	0.0160/0.0128/0.0020
	SapBERT-from-PubMedBERT-fulltext	**0.8828/0.8958/0.0071**	**0.8878/0.8980/0.0059†**	**0.8828/0.8960/0.0072**	**0.8666/0.8787/0.0068**	**0.0352/0.0317/0.0019**	**0.7903/0.8113/0.0113**	**0.8020/0.8191/0.0095**	**0.0152/0.0135/0.0014**
	BiomedNLP-BiomedBERT-abstract	0.8791/0.8926/0.0069	0.8855/0.8993/0.0076†	0.8788/0.8930/0.0071	0.8639/0.8792/0.0083	0.0365/0.0327/0.0020	0.7843/0.8060/0.0111	0.7987/0.8209/0.0125	0.0159/0.0137/0.0016
	BiomedNLP-BiomedBERT-abstract-fulltext	0.8802/0.8891/0.0078	0.8855/0.8958/0.0079†	0.8797/0.8895/0.0083	0.8642/0.8755/0.0084	0.0359/0.0333/0.0020	0.7860/0.8004/0.0123	0.7984/0.8137/0.0122	0.0165/0.0143/0.0015
SciBERT	Baseline Model	0.8542/0.8863/0.0228	0.8592/0.8983/0.0239	0.8522/0.8854/0.0244	0.8360/0.8778/0.0249	0.0424/0.0343/0.0055	0.7461/0.7958/0.0346	0.7599/0.8214/0.0346	0.0181/0.0134/0.0051
	all-MiniLM-L12-v2	0.8637/0.8830/0.0164	0.8700/0.8893/0.0178	0.8631/0.8832/0.0168	0.8382/0.8580/0.0209	0.0406/0.0352/0.0041	0.7604/0.7906/0.0251	0.7750/0.8044/0.0252	0.0175/0.0143/0.0030
	BioBERT-base-cased-v1.2	0.8611/0.8833/0.0226	0.8649/0.8914/0.0253	0.8599/0.8830/0.0247	0.8334/0.8624/0.0273	0.0415/0.0356/0.0057	0.7567/0.7910/0.0340	0.7682/0.8079/0.0341	0.0170/0.0131/0.0034
	SapBERT-from-PubMedBERT-fulltext	**0.8667/0.8799/0.0111**	**0.8724/0.8898/0.0143**	**0.8658/0.8799/0.0113**	**0.8379/0.8634/0.0158**	**0.0397/0.0368/0.0024**	**0.7649/0.7856/0.0171**	**0.7782/0.8054/0.0213**	**0.0168/0.0136/0.0040**
	BiomedNLP-BiomedBERT-abstract	0.8644/0.8772/0.0130	0.8688/0.8876/0.0155	0.8634/0.8771/0.0140	0.8388/0.8493/0.0112	0.0405/0.0371/0.0032	0.7614/0.7813/0.0198	0.7725/0.8015/0.0234	0.0157/0.0134/0.0022
	BiomedNLP-BiomedBERT-abstract-fulltext	0.8657/0.8789/0.0129	0.8721/0.8902/0.0131	0.8651/0.8796/0.0130	0.8374/0.8543/0.0116	0.0402/0.0368/0.0034	0.7634/0.7840/0.0198	0.7778/0.8056/0.0203	0.0157/0.0135/0.0029

Note: Bold values indicate the configurations showing the best overall performance within each classifier group, particularly considering Macro-F1. †Indicates statistical significance ($p < 0.05$) compared to the baseline model within the same classifier group.

In experimental studies, to analyze the contribution of general-purpose contextual representation to CombineTM, the all-MiniLM-L12-v2 model, introduced based on the SBERT approach, was employed. Examining the results presented in Table 7, it shows that the topic probability distributions obtained with this model significantly improve MLTC performance, especially on general-purpose classification models such as BERT (google-bert/bert-base-cased) (0.7604 → 0.8405 Macro-F1) and DistilBERT (distilbert/distilbert-base-cased) (0.8149 → 0.8413 Macro-F1). The p-values presented in Table 8 are computed using a paired t-test across multiple runs between the baseline models and the best-performing embedding models. The results indicate that the observed improvements are statistically significant for the BERT and DistilBERT models, with p-values of 0.0029 and 0.0104, respectively. In domain-adaptive models, a limited but consistent performance increase was achieved in domain-adaptive models. Statistically significant improvements were obtained for ClinicalBERT and BioBERT-v1.1. A particularly noteworthy performance improvement was observed in the ClinicalBERT model. This model, trained on clinical texts, performed better on a HoC dataset constructed from scientific article abstracts in the cancer literature when representations enriched with topic probability distributions were used (0.8066 → 0.8461 Macro-F1). The findings demonstrate that, in addition to improvements observed in general-purpose models, representations enhanced with topic probability distributions provide a complementary representation for this model, which is specialized in a different biomedical subdomain. This suggests that the difference between the dataset on which the model was pre-trained and the target dataset can be balanced by the conceptual knowledge provided at the topic level. In this context, SBERT-based general-purpose representations play a complementary role, especially for models with limited domain knowledge, while also providing a supportive and enriching representation signal for domain-adaptive models.

Table 8: Statistical significance (p-values) of the best performing embedding models compared to the baseline.

Classifier	Embedding Model	p-Value
BERT	all-MiniLM-L12-v2	0.0029
	BiomedNLP-BiomedBERT-base-uncased-abstract	0.0018
DistilBERT	all-MiniLM-L12-v2	0.0104
	BioBERT-base-cased-v1.2	0.0346
BioMed-RoBERTa	BiomedNLP-BiomedBERT-base-uncased-abstract	0.0133
ClinicalBERT	BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext	0.0001
PubMedBERT	BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext	0.0397
BioBERT-v1.1	SapBERT-from-PubMedBERT-fulltext	0.0008
SciBERT	SapBERT-from-PubMedBERT-fulltext	0.0669

The SBERT architecture, specifically the all-MiniLM-L12-v2 model, transforms documents into dense vector representations that provide a semantic signal to the contextual component of CombinedTM. However, considering the domain-specific terminological depth of the biomedical literature, general-purpose SBERT representations can be limited, particularly in distinguishing domain-specific biomedical terms and subtle conceptual distinctions. To overcome this limitation, instead of commonly used general-purpose SBERT variants, the contribution of domain-adaptive language models enriched with biomedical corpora, such as BioBERT, SapBERT, and PubMedBERT, was examined. These models were employed within the

SBERT framework, which provides document-level contextual embeddings by aggregating token-level representations via a mean pooling strategy. Thus, in the contextual component of CombinedTM, representations produced by both general-purpose and biomedical-specific language models were examined comparatively. Furthermore, the impact of topic probability distributions on MLTC performance was analyzed, providing insight into the effectiveness of these models in learning such distributions.

The findings presented in Table 7 reveal that topic probability distributions obtained via topic modeling with domain-adaptive embeddings improved the MLTC performance of both general-purpose and domain-adaptive PLMs. While the general-purpose models, BERT and DistilBERT, exhibit limited performance when fine-tuned solely on contextual representations, MLTC performance improved when topic probability distributions obtained using the proposed embedding models were incorporated as additional representations during fine-tuning. The results particularly demonstrate that general-purpose models benefit strongly from these additional representations. The most significant improvements in MLTC performance were observed under different topic modeling configurations for each classifier group. For the BERT model, the best MLTC performance was achieved when topic probability distributions were derived using the BiomedNLP–PubMedBERT-abstract embedding model. Using representations enhanced with topic probability distributions obtained from CombinedTM, which was trained with this embedding, the Macro-F1 score of the fine-tuned BERT model increased from 0.7604 to 0.8458. In the DistilBERT model, significant improvements were observed with BioBERT-base-cased-v1.2 configuration, where representations enriched with topic probability distributions increased the Macro-F1 score of the baseline model from 0.8149 to 0.8347. The p-values presented in Table 8 reveal that the performance improvements observed in general-purpose models are statistically significant.

Domain-adaptive PLMs already exhibit strong baseline performance on MLTC tasks due to their training on large-scale corpora of biomedical and scientific texts using various pre-training strategies. Nevertheless, the experimental findings presented in Table 7 demonstrate that the topic probability distributions obtained via CombinedTM using domain-adaptive embedding models provide additional text representations that further improve the MLTC performance of domain-adaptive PLMs. In particular, BioMed-RoBERTa and ClinicalBERT models, fine-tuned using representations enriched with topic probability distributions, achieved significant improvements in evaluation metrics on the test dataset compared to baseline models. In the configuration where BiomedNLP-PubMedBERT-abstract embeddings were used for topic modeling, the Macro-F1 score of BioMed-RoBERTa increased from 0.8665 to 0.8819. Similarly, the Macro-F1 score of ClinicalBERT increased from 0.8066 to 0.8528 when BiomedNLP-PubMedBERT-abstract-fulltext embeddings were used. These improvements demonstrate that, despite the strong contextual representations of domain-adaptive models, presenting the topic of the document as an additional representation further enhances MLTC performance. On the other hand, the configuration using BiomedNLP-PubMedBERT-abstract-fulltext embedding for topic modeling and PubMedBERT for classification achieved the highest performance across all experimental evaluations, with a Macro-F1 score of 0.8900. This result highlights that combining domain-adaptive contextual representations with higher-level conceptual knowledge obtained from topic modeling provides an effective representation strategy for biomedical MLTC tasks.

To evaluate the robustness of the proposed approach, a sensitivity analysis was conducted on BERT and PubMedBERT, which achieved the highest classification performance among the general-purpose and domain-adaptive pre-trained language models, respectively. Different initial values of K (10, 15, and 20) and projection dimensions were systematically evaluated, as presented in Figs. 3 and 4. The best-performing configurations were selected based on the mean Macro-F1 scores over 10 runs (Table 7), and sensitivity analysis was conducted on representative runs (seeds) corresponding to the maximum Macro-F1 values

reported in the table for each configuration. As shown in Fig. 3, the highest Macro-F1 score (0.9026) for the PubMedBERT + BiomedNLP-PubMedBERT-abstract-fulltext configuration was achieved with the combination of $K = 15$ topics and a projection dimension of 128. This result indicates a configuration in which domain-specific embeddings and topic probability distributions are most effectively aligned. In contrast, for the BERT + BiomedNLP-PubMedBERT-abstract configuration presented in Fig. 4, it was observed that the $K = 14$ (initial K = 15) scenario exhibited a significant performance decrease in the 256-dimensional projection. This indicates that even under the best-performing operation, the model model may exhibit sensitivity in certain hyperparameter regions. However, a common trend in both models is that the K = 15 and a projection dimension of 128 provides more dense and discriminative feature representations, resulting in more stable and higher performance.

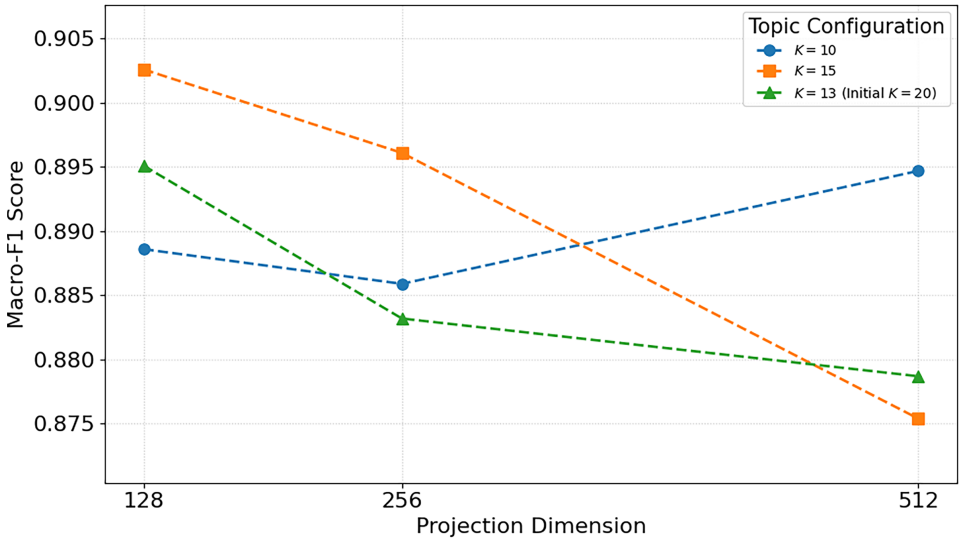

Figure 3: Impact of topic count (K) and projection dimension on PubmedBERT+ BiomedNLP-PubMedBERT-abstract-fulltext configuration.

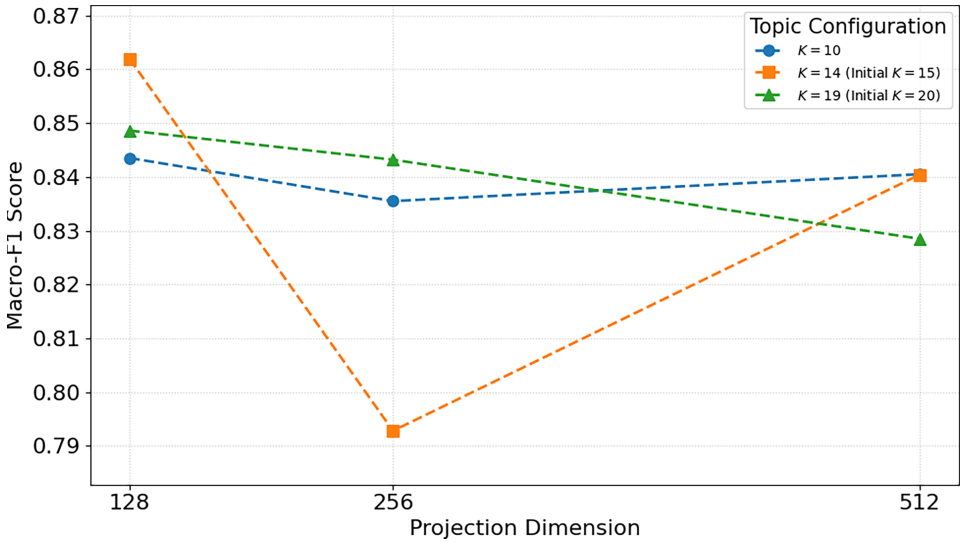

Figure 4: Impact of topic count (K) and projection dimension on BERT+ BiomedNLP-PubMedBERT-abstract configuration.

To comprehensively evaluate the operational efficiency and computational cost of the proposed methodology, the time consumption of each component of the system was analyzed in detail. GPU acceleration was used for computationally intensive stages such as contextual embedding generation and topic model training and inference, while sequential and lower-cost operations like text preprocessing and post-processing were executed on the CPU. The preprocessing steps required for constructing the BoW based statistical representation, as described in Section Data Cleaning, take on average 12.21 s. The most significant difference in computational cost is observed in the contextual embedding generation phase. While the general-purpose model all-MiniLM-L12-v2 completes this process in approximately 3.94 s, this time increases to an average of 39.45 s for domain-adaptive embedding models. In contrast, topic model training exhibits a relatively stable cost, taking on average 18.60 s. The inference time, supporting the practical applicability of the system performance, is quite low at approximately 0.35 s. The final stage of the process, post-processing, which reduces semantic noise, has a negligible computational cost. In terms of total time, general-purpose models such as all-MiniLM-L12-v2 complete the entire process in approximately 35.73 s, while this time increases to average of 73.74 s for domain-adaptive embedding models. This difference can be considered a reasonable computational trade-off given the richer semantic representations and the observed improvements in classification performance. Furthermore, compared with the baseline PLMs, integrating topic information into the PLM representations did not lead to any substantial increase in fine-tuning or inference time. BERT-based models exhibited fine-tuning times of approximately 355–410 s, total inference times of 9.1–11.1 s, and per-sample inference times of 0.029–0.035 s. For DistilBERT and ClinicalBERT, fine-tuning times were approximately 180–232 s, with total inference times ranging from 4.4 to 5.6 s and per-sample inference times between 0.014 and 0.018 s. Similarly, the domain-specific BioMed-RoBERTa, PubMedBERT, BioBERT, and SciBERT models required approximately 349–414 s for fine-tuning, 8.6–11.3 s for total inference, and 0.027–0.036 s per sample. Overall, the topic-enhanced models exhibited fine-tuning and inference times that were virtually identical to those of their corresponding baseline models. The comparable runtime ranges confirm that the integration of topic information does not introduce a systematic computational overhead in terms of model parameters or computational complexity. The minor runtime variations observed are more likely attributable to differences in random seeds and transient variations associated with the training process and hardware environment rather than the topic integration itself. These findings indicate that the primary computational cost of the proposed framework arises not from fine-tuning the PLMs, but from the offline topic representation extraction stage. As reported in Table 9, the additional one-time cost incurred during topic representation extraction can be regarded as an acceptable computational trade-off considering the richer semantic representations obtained and the substantial improvements achieved in classification performance.

Considering the multi-label and imbalanced nature of the HoC dataset used in the experimental studies, the Macro-F1 score was emphasized for evaluating MLTC performance. Furthermore, the reported Micro-F1, Weighted-F1, Hamming Loss, Jaccard Micro/Macro, and Label Ranking Loss metrics enabled a more comprehensive analysis of model behavior from different perspectives. Experimental results that are presented in Table 7 show that, compared to the fine-tuned baseline models, the models fine-tuned with topic-enriched representations exhibited improvements in Micro-F1 and Weighted-F1 scores, indicating that overall predictive performance on dominant classes in the dataset was preserved, with the accompanying improvements in MCC providing additional evidence of improved prediction quality. Meanwhile, the decrease in Hamming Loss values and the improvements observed in Jaccard metrics indicate reduced label-level prediction errors and increased overlap between the predicted and ground-truth label sets. Ranking Loss, especially in multi-label scenarios, provides a critical metric for evaluating the model's ability to prioritize relevant labels correctly. When all metric results presented in Table 7 are evaluated comparatively,

consistent improvements are observed not only in Macro-F1 scores but also in all other metrics. This indicates that incorporating topic information has a positive impact not only on rare label prediction but also on overall classification behavior and label ranking performance.

Table 9: Runtime analysis of the proposed topic representation generation pipeline.

Processing Stage	Runtime (s)
Text preprocessing	12.21
Contextual embedding generation (general-purpose model)	3.94
Contextual embedding generation (domain-adaptive models, avg.)	39.45
Topic model training	18.60
Topic inference	0.35
Total runtime (general-purpose configuration)	35.73
Total runtime (domain-adaptive configurations, avg.)	73.74

To further contextualize the proposed approach, previous studies employing the HoC dataset were reviewed. Studies investigated PLM-based transformer architectures [6,10], hybrid approaches [34], LLM-based ensemble approaches [12], and machine learning methods [13]. However, some of these studies adopted different sampling strategies, data splits, and evaluation protocols, making their reported results not directly comparable Table 10. Therefore, quantitative comparisons were limited to studies using the same HoC benchmark split. Under this setting, the proposed PubMedBERT + CombinedTM model achieved a Macro-F1 score of 0.8900, demonstrating competitive performance compared with previous studies using the same data partition. Nevertheless, the primary contribution of this study is not merely achieving the highest performance score, but demonstrating that integrating CombinedTM-derived topic probability distributions with contextual representations consistently improves performance, particularly for general-purpose pre-trained language models. These findings suggest that the proposed text representation approach has the potential to improve the competitiveness of general-purpose PLMs in biomedical multi-label text classification tasks.

Table 10: Comparison of recent studies on the HoC dataset and the proposed approach.

Study	Models	Approach	HoC Train/Dev/Test Split	Reported Performance
[6]	MLP, BiLSTM, and transformers	Evaluation of TF-IDF, embedding-based, and transformer-based text representations	80/20 train (10% val.)/test	Micro-F1 = 0.8300 (PubMedBERT)
[10]	BioBERT	Shared transformer backbone with label-specific and label-pair representations	1108/157/315	Macro-F1 = 0.8733 (mean)

(Continued)

Table 10 (continued)

Study	Models	Approach	HoC Train/Dev/Test Split	Reported Performance
[12]	GPT-4o, BERT, PEGASUS, and BART	Multi-LLM pipeline with ensemble learning	Stratified samples (300/500/1000)	Macro-F1 = 0.7811
[13]	SVM	BioWordVec and MeSH embeddings	30-fold CV	Average class-wise F1 = 98.95%
[34]	PubMedBERT and CNN	PubMedBERT–CNN hybrid approach	80/20 train (10% val.)/test	F1 = 0.8900
Our Study	General-purpose and domain-adaptive PLMs	Fusion of CombinedTM topic probabilities with contextual PLM representations	1108/157/315	Macro-F1 = 0.8900 BERT: 0.7604 → 0.8458

Overall, the experimental findings indicate that text representations enriched with topic probability distributions have a positive impact on MLTC performance. These representations improve average performance across all model groups and yield consistent results across multiple runs. Furthermore, as shown in Table 7, the relatively low standard deviations observed across the 10 independent runs (e.g., a Macro-F1 standard deviation of 0.0079 for the best PubMedBERT configuration) indicate low performance variance and demonstrate the robustness of the proposed enhanced representation strategy. Moreover, although the magnitude of the performance improvements varies across different PLMs, incorporating topic probability distributions consistently maintained or improved the performance of the corresponding baseline models, indicating that the proposed enhanced representation provides complementary rather than detrimental information. The p-values obtained using a paired t-test further demonstrate that the observed performance gains are statistically significant. To examine the thematic disentanglement capabilities provided by this model in more detail, the distribution of documents obtained using the BiomedNLP–PubMedBERT-abstract-fulltext embedding model, which yields the best performance in combination with PubMedBERT, is presented in Table 11, and its visual representation is shown in Fig. 5. In addition, the conceptual structure of the extracted topic space is illustrated by listing the top 10 keywords learned by the model for each topic.

Consequently, when fine-tuned with the enhanced representations, the BERT model achieved Micro/Macro/Weighted-F1 scores of 0.8469/0.8458/0.8445 on the test dataset, while also exhibiting the best performance in terms of Hamming Loss of 0.0450, Jaccard Micro/Macro of 0.7345/0.7397, and Label Ranking Loss of 0.0243. Although the general-purpose embedding model all-MiniLM-L12-v2 used in the representation strategy did not contribute as highly to the BERT model as domain-adaptive embeddings, it achieved a performance level close to domain-adaptive models in terms of Micro/Macro/Weighted-F1 scores of 0.8384/0.8405/0.8351, Hamming Loss of 0.0468, Jaccard Micro/Macro of 0.7224/0.7324, and Label Ranking Loss of 0.0254. These results are noteworthy because they demonstrate that the proposed topic-based representation strategy significantly enhances the classification capacity of general-purpose models across both stages. An examination of domain-adaptive models reveals that the proposed representation strategy consistently provides a complementary contribution. In this context, the configuration in which the BiomedNLP-PubMedBERT-abstract-fulltext embedding was employed within the contextual component

of CombinedTM and PubMedBERT was used as the classifier, which achieved the highest performance, with Micro/Macro/Weighted-F1 scores of 0.8839/0.8900/0.8838. This configuration also outperformed others across all evaluation metrics, including a Hamming Loss of 0.0353, Jaccard Micro/Macro scores of 0.7921/0.8063, and a Label Ranking Loss of 0.0130. The experimental findings from this study demonstrate that integrating document-level topic probability distributions as an additional representation strategy into the biomedical multi-label text classification process yields consistent improvements in overall classification performance. In particular, the gains are most pronounced for general-purpose models.

Table 11: Document-topic distribution from BiomedNLP-PubMedBERT-abstract-fulltext.

No	Topic	Train Set	Val. Set	Test Set	Total
Topic 1	high tissue case positive low normal patient expression Ki-67 stage	47	7	17	71
Topic 2	HCC cytometry flow phase assay progression EMT apoptosis transfected could	43	4	8	55
Topic 3	pathway inhibitor signal Akt activation activity promote death downregulation glucose	49	4	7	60
Topic 4	line vitro vivo reduce apoptosis $p < 0.05$ proliferation compare expression breast	75	6	17	98
Topic 5	Genotoxic (-1)L hepatotrophic IL-6-arginase Neuro-2a receptor-pore OXPHOS-related [35S]methionine gene-coding Northeast	76	15	15	106
Topic 6	Bcl-2 ATP effect pathway phase receptor antagonist LNCaP functional cell	99	8	36	143
Topic 7	sequence nucleotide chromosomal ion chromosome DNA aberration break deletion lesion	77	9	21	107
Topic 8	therapeutic endothelial compound glioblastoma stem strategy reverse antiproliferative data culture	179	28	65	272
Topic 9	patient month year survival metastasis case recurrence node risk lymph	70	17	16	103
Topic 10	DNA antioxidant damage oxidative different protect genetic repair alteration BaP	108	14	38	160
Topic 11	treatment mouse vivo immune tumor reduce effect vitro melanoma model	19	2	2	23
Topic 12	growth breast proliferation vivo vitro EGFR cell xenograft progression line	30	8	12	50
Topic 13	upon ectopic p53 regulates transcriptional CDK regulation checkpoint Together kinase	50	7	17	74

(Continued)

Table 11 (continued)

No	Topic	Train Set	Val. Set	Test Set	Total
Topic 14	inflammation immune inflammatory mouse cytokine diet response liver animal oxidative	64	13	20	97
Topic 15	cellular poorly regulate contact oxygen regulates metabolic know CDK understood	122	15	24	161

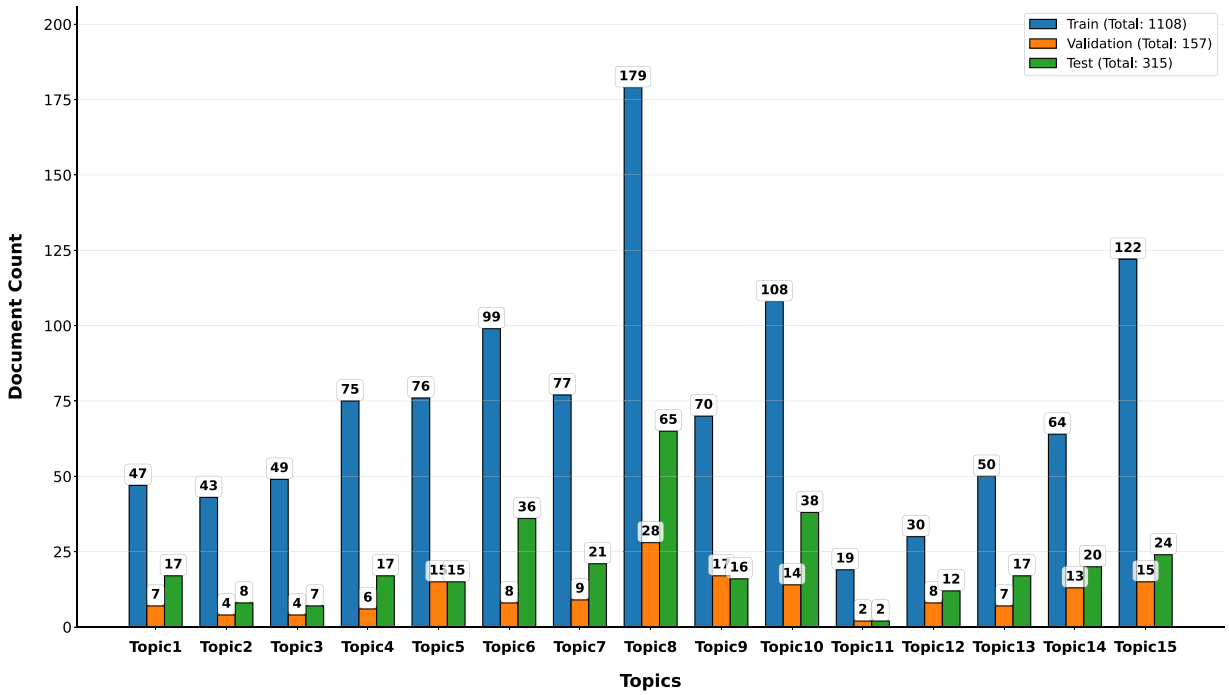

Figure 5: Document-topic distribution across all datasets from BiomedNLP-PubMedBERT-abstract-fulltext.

6 Limitations and Research Opportunities

This study evaluates the potential of the proposed topic-enhanced models in biomedical multi-label text classification tasks on the HoC dataset. The findings reveal that text representations enriched with topic probability distributions yield improved performance on this dataset. However, the this study is limited to the HoC dataset which is structured with 10 different hallmark labels specific to the cancer literature. The HoC dataset addresses the multi-label classification problem in the biomedical literature through a hallmark based conceptual framework. It is observed that numerous studies in the biomedical NLP literature focus on the this dataset for MLTC task [6,10,12,13,34]. To the best of our knowledge, the lack of alternative open-source datasets in the existing literature, structured with 10 hallmark labels, multi-labeled, and document-level labeled, focusing on the cancer literature, limits the direct comparative evaluation of the proposed approach on different datasets under the same problem definition. While this constitutes a constraint on the scope of the study, also highlights the significance of the contribution achieved by the proposed method within this specific problem setting. The proposed method was evaluated using numerous domain-specific and general-purpose models, both in the contextual text representation phases of CombinedTM and text

classification phase. A consistent improvement was observed in experiments conducted with 10 repetitions. The consistent improvement achieved across different language models with varying architectures and pre-training backgrounds supports the effectiveness of the proposed method. In this regard, evaluating the proposed approach on benchmark biomedical datasets (e.g., LitCovid and MIMIC-III), as well as on clinical notes or biomedical texts with different thematic focuses, presents significant potential for future studies.

7 Conclusion

This study presents an integrated approach that incorporates topic modeling into the classification process to improve biomedical MLTC. The proposed method aims to develop a robust text representation strategy for classification by leveraging the strong contextual representation capabilities of PLMs and supporting them with context-aware topic probability distributions obtained from CombinedTM. The findings indicate that using topic information as an additional representational signal can provide more consistent and discriminative classification decisions. In this respect, the study offers a complementary perspective to the relevant literature by systematically investigating the contribution of topic probability distribution to multi-label classification performance in a biomedical domain.

Comprehensive experiments conducted on the HoC dataset, which constitutes an important data source in the cancer literature, demonstrate that a representation enhanced with topic probability distributions yields meaningful performance improvements for both general-purpose and biomedical domain-adapted language models. In particular, the findings show that a general-purpose model such as BERT exhibits a substantial improvement when supported by topic-aware text representations (Macro F1: 0.7604 → 0.8458). These results suggest that general-purpose models can achieve competitive performance on biomedical tasks without requiring costly domain-adaptation processes. At the same time, the observed performance gains for domain-adaptive models with this additional representation method (BioMedRoBERTa: 0.8665 → 0.8819 Macro-F1; PubMedBERT + CombinedTM = 0.8900 Macro-F1) indicate that the proposed approach provides a model-agnostic, generalizable contribution. In future work, the proposed approach can be applied to different biomedical text sources, such as literature, EHR, and clinical reports, to more comprehensively evaluate its generalizability and scalability. Additionally, the performance of large language models (LLMs) on biomedical MLTC will be investigated.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization: Oyku Berfin Mercan, Nezihe Turhan Turan and Aytuğ Onan; Methodology: Oyku Berfin Mercan, Nezihe Turhan Turan and Aytuğ Onan; Writing—original draft preparation: Oyku Berfin Mercan; Writing—review and editing: Nezihe Turhan Turan and Aytuğ Onan; Supervision: Nezihe Turhan Turan and Aytuğ Onan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: A publicly available open-source Hallmarks of Cancer (HoC) dataset is used in this study. Details of the dataset are provided in Section 4.1.

Ethics Approval: Not applicable.

Conflicts of Interest: Given his role as Editorial Board Member of this journal, Aytuğ Onan had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

References

1. Kesiku CY, Chaves-Villota A, Garcia-Zapirain B. Natural language processing techniques for text classification of biomedical documents: a systematic review. Information. 2022;13(10):499. doi:10.3390/info13100499.
2. Katwe PK, Khamparia A, Gupta D, Dutta AK. Methodical systematic review of abstractive summarization and natural language processing models for biomedical health informatics: approaches, metrics and challenges. ACM Trans Asian Low Resour Lang Inf Process. 2023;26(4):3600230. doi:10.1145/3600230.
3. Eguia H, Sánchez-Bocanegra CL, Vinciarelli F, Alvarez-Lopez F, Saigí-Rubió F. Clinical decision support and natural language processing in medicine: systematic literature review. J Med Internet Res. 2024;26(2):e55315. doi:10.2196/55315.
4. Houssein EH, Mohamed RE, Ali AA. Machine learning techniques for biomedical natural language processing: a comprehensive review. IEEE Access. 2021;9:140628–53. doi:10.1109/access.2021.3119621.
5. Parwez MA, Fazil M, Arif M, Nafis MT, Auwul MR. Biomedical text classification using augmented word representation based on distributional and relational contexts. Comput Intell Neurosci. 2023;2023(1):2989791. doi:10.1155/2023/2989791.
6. Kumawat H, Sharan A, Verma S. Impact analysis of text representation on biomedical multi-label text classification with deep learning. Procedia Comput Sci. 2025;258(7):3294–304. doi:10.1016/j.procs.2025.04.587.
7. Hassan E, Abd El-Hafeez T, Shams MY. Optimizing classification of diseases through language model analysis of symptoms. Sci Rep. 2024;14(1):1507. doi:10.1038/s41598-024-51615-5.
8. Chaichulee S, Promchai C, Kaewkomon T, Kongkamol C, Ingviya T, Sangsupawanich P. Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language processing. PLoS One. 2022;17(8):e0270595. doi:10.1371/journal.pone.0270595.
9. Liu L, Perez-Concha O, Nguyen A, Bennett V, Jorm L. Automated ICD coding using extreme multi-label long text transformer-based models. Artif Intell Med. 2023;144(2):102662. doi:10.1016/j.artmed.2023.102662.
10. Chen Q, Du J, Allot A, Lu Z. LitMC-BERT: transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation. IEEE/ACM Trans Comput Biol Bioinform. 2022;19(5):2584–95. doi:10.1109/TCBB.2022.3173562.
11. He Y, Xiong Q, Ke C, Wang Y, Yang Z, Yi H, et al. MCICT: graph convolutional network-based end-to-end model for multi-label classification of imbalanced clinical text. Biomed Signal Process Control. 2024;91(1):105873. doi:10.1016/j.bspc.2023.105873.
12. Sakai H, Lam SS. QUAD-LLM-MLTC: large language models ensemble learning for healthcare text multi-label classification. arXiv:2502.14189. 2025. doi:10.48550/arXiv.2502.14189.
13. Verma S, Sharan A, Malik N. Efficient classification of hallmark of cancer using embedding-based support vector machine for multilabel text. New Gener Comput. 2024;42(4):685–714. doi:10.1007/s00354-024-00248-3.
14. Yogarajan V, Pfahringer B, Smith T, Montiel J. Improving predictions of tail-end labels using concatenated BioMed-transformers for long medical documents. arXiv:2112.01718. 2021. doi:10.48550/arXiv.2112.01718.
15. Yogarajan V, Montiel J, Smith T, Pfahringer B. Seeing the whole patient: using multi-label medical text classification techniques to enhance predictions of medical codes. arXiv:2004.00430. 2020. doi:10.48550/arXiv.2004.00430.
16. Kalyan KS, Sangeetha S. SECNLP: a survey of embeddings in clinical natural language processing. J Biomed Inform. 2020;101:103323. doi:10.1016/j.jbi.2019.103323.
17. Davagdorj K, Wang L, Li M, Pham VH, Ryu KH, Theera-Umpon N. Discovering thematically coherent biomedical documents using contextualized bidirectional encoder representations from transformers-based clustering. Int J Environ Res Public Health. 2022;19(10):5893. doi:10.3390/ijerph19105893.
18. Saad E, Kishk S, Ali-Eldin A, Saleh AI. The effect of transformer based pre-trained language models on biomedical relation extractions. Mansoura Eng J. 2025;50(4):14. doi:10.58491/2735-4202.3304.
19. Wang X, Liu G, Zhu B, He J, Zheng H, Zhang H. Pre-trained language models and few-shot learning for medical entity extraction. In: Proceedings of the 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA); 2025 May 23–25; Changsha, China. p. 1243–7.

20. Doneva SE, Qin S, Sick B, Ellendorff T, Goldman JP, Schneider G, et al. Large language models to process, analyze, and synthesize biomedical texts: a scoping review. Discover Artif Intell. 2024;4(1):107. doi:10.1007/s44163-024-00197-2.
21. Madan S, Lentzen M, Brandt J, Rueckert D, Hofmann-Apitius M, Fröhlich H. Transformer models in biomedicine. BMC Med Inform Decis Mak. 2024;24(1):214. doi:10.1186/s12911-024-02600-5.
22. Aamir N, Raza A, Iqbal MW, Hamid K, Nazir Z, Asif A, et al. Topic modeling empowered by a deep learning framework integrating BERTopic, XLM-R, and GPT. J Comput Biomed Inform. 2025;8(2):1–18. doi:10.21015/vtcs.v13i2.2144.
23. Hidayat AA, Nirwantono R, Budiarto A, Pardamean B. BERT-based topic modeling approach for malaria research publication. In: Proceedings of the 2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS); 2022 Nov 16–17; Jakarta, Indonesia. p. 326–31. doi:10.1109/icimcis56303.2022.10017743.
24. Galli C, Cusano C, Meleti M, Donos N, Calciolari E. Topic modeling for faster literature screening using transformer-based embeddings. Metrics. 2024;1(1):2. doi:10.3390/metrics1010002.
25. Veeranki SPK, Abdulnazar A, Kramer D, Kreuzthaler M, Lumenta DB. Multi-label text classification via secondary use of large clinical real-world data sets. Sci Rep. 2024;14(1):26972. doi:10.1038/s41598-024-76424-8.
26. Yogarajan V, Montiel J, Smith T, Pfahringer B. Transformers for multi-label classification of medical text: an empirical comparison. In: Artificial intelligence in medicine. Cham, Switzerland: Springer International Publishing; 2021. p. 114–23. doi:10.1007/978-3-030-77211-6_12.
27. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794. 2022. doi:10.48550/arXiv.2203.05794.
28. Bianchi F, Terragni S, Hovy D. Pre-training is a hot topic: contextualized document embeddings improve topic coherence. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers); 2021 Aug 1–6; Online. Stroudsburg, PA, USA: Association for Computational Linguistics; 2021. p. 759–66.
29. Lenivtceva I, Slasten E, Kashina M, Kopanitsa G. Applicability of machine learning methods to multi-label medical text classification. In: Computational Science—ICCS 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 509–22. doi:10.1007/978-3-030-50423-6_38.
30. Zhang Y, Li X, Liu Y, Li A, Yang X, Tang X. A multilabel text classifier of cancer literature at the publication level: methods study of medical text classification. JMIR Med Inform. 2023;11:e44892. doi:10.2196/44892.
31. Kementchedjhieva Y, Chalkidis I. An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text. arXiv:2305.05627. 2023. doi:10.48550/arXiv.2305.05627.
32. ul haq Inaam M, Li Q, Mahmood K, Shafique A, Ullah R. BioElectra-BiLSTM-dual attention classifier for optimizing multilabel scientific literature classification. Comput J. 2025;68(5):565–76. doi:10.1093/comjnl/bxae132.
33. Lin SJ, Yeh WC, Chiu YW, Chang YC, Hsu MH, Chen YS, et al. A BERT-based ensemble learning approach for the BioCreative VII challenges: full-text chemical identification and multi-label classification in PubMed articles. Database. 2022;2022:baac056. doi:10.1093/database/baac056.
34. El-Azizy Z, Aouragh SL, Ouatik El Alaoui S. Multi-label biomedical text classification of abstracts according to the hallmarks of cancer. In: Proceedings of the 2025 International Conference on Intelligent Systems: Theories and Applications (SITA); 2025 Oct 20–21; Rabat, Morocco. p. 1–6. doi:10.1109/sita67914.2025.11273579.
35. Lebeña N, Blanco A, Pérez A, Casillas A. Preliminary exploration of topic modelling representations for electronic health records coding according to the international classification of diseases in Spanish. Expert Syst Appl. 2022;204(4):117303. doi:10.1016/j.eswa.2022.117303.
36. Pérez J, Pérez A, Casillas A, Gojenola K. Cardiology record multi-label classification using latent Dirichlet allocation. Comput Methods Programs Biomed. 2018;164(1):111–9. doi:10.1016/j.cmpb.2018.07.002.
37. Mortezaagha P, Makarand Joshi A, Rahgozar A. Inconsistency detection in cancer data classification using explainable-AI. BMC Artif Intell. 2025;1(1):5. doi:10.1186/s44398-025-00005-6.
38. Rijcken E, Kaymak U, Scheepers F, Mosteiro P, Zervanou K, Spruit M. Topic modeling for interpretable text classification from EHRs. Front Big Data. 2022;5:846930. doi:10.3389/fdata.2022.846930.

39. Voskergian D, Jayousi R, Yousef M. Topic selection for text classification using ensemble topic modeling with grouping, scoring, and modeling approach. Sci Rep. 2024;14(1):23516. doi:10.1038/s41598-024-74022-2.
40. Luk S. LLM finetuning for multilabel classification. 2024 [cited 2026 Jan 1]. Available from: https://github.com/sirluk/llm$\delimiter"0383383$_finetuning/tree/main/llm$\delimiter"0383383$_multilabel_clf.
41. Baker S, Silins I, Guo Y, Ali I, Högberg J, Stenius U, et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer. Bioinformatics. 2016;32(3):432–40. doi:10.1093/bioinformatics/btv585.
42. Chen Q, Hu Y, Peng X, Xie Q, Jin Q, Gilson A, et al. A systematic evaluation of large language models for biomedical natural language processing: benchmarks, baselines, and recommendations. Zenodo. 2024. doi:10.5281/zenodo.14025500.
43. Yogarajan V, Gouk H, Smith T, Mayo M, Pfahringer B. Comparing high dimensional word embeddings trained on medical text to bag-of-words for predicting medical codes. In: Intelligent information and database systems. Cham, Switzerland: Springer International Publishing; 2020. p. 97–108. doi:10.1007/978-3-030-41964-6_9.
44. Galli C, Donos N, Calciolari E. Performance of 4 pre-trained sentence transformer models in the semantic query of a systematic review dataset on peri-implantitis. Information. 2024;15(2):68. doi:10.3390/info15020068.
45. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40. doi:10.1093/bioinformatics/btz682.
46. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2021 Jun 6–11; Online. Stroudsburg, PA, USA: Association for Computational Linguistics; 2021. p. 4228–38.
47. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. 2022;3(1):1–23. doi:10.1145/3458754.
48. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2019. doi:10.48550/arXiv.1810.04805.
49. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. 2020. doi:10.48550/arXiv.1910.01108.
50. Ren L, Belkadi S, Han L, Del-Pinto W, Nenadic G. Synthetic4Health: generating annotated synthetic clinical letters. Front Digit Health. 2025;7:1497130. doi:10.3389/fdgth.2025.1497130.
51. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. arXiv:1903.10676. 2019. doi:10.48550/arXiv.1903.10676.
52. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't stop pretraining: adapt language models to domains and tasks. arXiv:2004.10964. 2020. doi:10.48550/arXiv.2004.10964.
53. Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min. 2021;14(1):13. doi:10.1186/s13040-021-00244-z.
54. Tamura J, Itaya Y, Hayashi K, Yamamoto K. Statistical inference of the Matthews correlation coefficient for multiclass classification. J Appl Stat. 2026:1–26. doi:10.1080/02664763.2026.2666305.