applied sciences

MDPI

Article

Multi-Task Topic Analysis Framework for Hallmarks of Cancer with Weak Supervision

Erdenebileg Batbaatar [1], Van-Huy Pham [2] and Keun Ho Ryu [2,3,*]

[1] Database and Bioinformatics Laboratory, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju 28644, Korea; erdenebileg11@gmail.com
[2] Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh 700000, Vietnam; phamvanhuy@tdtu.edu.vn
[3] Department of Computer Science, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju 28644, Korea
* Correspondence: khryu@tdtu.edu.vn or khryu@chungbuk.ac.kr; Tel.: +82-43-267-2254

Received: 31 December 2019; Accepted: 22 January 2020; Published: 24 January 2020

Abstract: The hallmarks of cancer represent an essential concept for discovering novel knowledge about cancer and for extracting the complexity of cancer. Due to the lack of topic analysis frameworks optimized specifically for cancer data, the studies on topic modeling in cancer research still have a strong challenge. Recently, deep learning (DL) based approaches were successfully employed to learn semantic and contextual information from scientific documents using word embeddings according to the hallmarks of cancer (HoC). However, those are only applicable to labeled data. There is a comparatively small number of documents that are labeled by experts. In the real world, there is a massive number of unlabeled documents that are available online. In this paper, we present a multi-task topic analysis (MTTA) framework to analyze cancer hallmark-specific topics from documents. The MTTA framework consists of three main subtasks: (1) cancer hallmark learning (CHL)—used to learn cancer hallmarks on existing labeled documents; (2) weak label propagation (WLP)—used to classify a large number of unlabeled documents with the pre-trained model in the CHL task; and (3) topic modeling (ToM)—used to discover topics for each hallmark category. In the CHL task, we employed a convolutional neural network (CNN) with pre-trained word embedding that represents semantic meanings obtained from an unlabeled large corpus. In the ToM task, we employed a latent topic model such as latent Dirichlet allocation (LDA) and probabilistic latent semantic analysis (PLSA) model to catch the semantic information learned by the CNN model for topic analysis. To evaluate the MTTA framework, we collected a large number of documents related to lung cancer in a case study. We also conducted a comprehensive performance evaluation for the MTTA framework, comparing it with several approaches.

Keywords: multi-task learning; topic analysis; semantic learning; convolutional neural network; latent semantic learning; biomedical domain; cancer hallmark; lung cancer

1. Introduction

Cancer is the second leading cause of death globally in 2018 and it is incredibly complicated [1]. The characteristics used to distinguish cancer cells from normal cells are called the hallmarks of cancer (HoC) [2,3]. The cancer hallmark is the fundamental principle of malignant transformation of a cancer cell and useful to understand tumor pathogenesis. It is becoming highly influential in cancer research [4–6]. There are common 10 hallmarks of normal cells required for malignant growth that have been proposed that provide an organizing principle to explain the variety of the biological processes leading to cancer [3]. These are sustaining proliferative signaling (SPS), evading growth suppressors

(EGSs), resisting cell death (RCD), enabling replicative immortality (ERI), inducing angiogenesis (IA), activating invasion and metastasis (AIM), genome instability and mutation (GIM), tumor-promoting inflammation (TPI), deregulating cellular energetics (DCE), and avoiding immune destruction (AID).

Most of the cancer research results have been published in the biomedical literature. The famous biomedical literature database is PubMed, which has indexed approximately 30 million citations including around 4 million cancer-related literature by 2019. As biomedical literature on servers grows exponentially, biomedical text mining has been intensively investigated to find information in a more accurate and efficient manner [7,8]. A large amount of biomedical literature in that database provides a great opportunity to extract useful knowledge for cancer research. Researchers use a keyword-based query to collect relevant literature from that database. Due to the complexity of cancer, a large number of keywords, their synonyms, and combinations are required, which is a very time-consuming task to retrieve relevant literature by using only a keyword. Most of the previous studies have used the comparatively small dataset annotated by the experts for the hallmarks of cancer. The dataset is small and the number of labeled documents, in particular, is very limited. To address this issue, we leverage a large number of documents related to lung cancer from PubMed in a semi-supervised manner.

Recently, deep learning (DL) based approaches have achieved state-of-the-art performance and are increasingly applied to most natural language processing (NLP) tasks [9–11]. NLP has benefited greatly from the convolutional neural network (CNN) [12] and recurrent neural network (RNN) [13] due to their high performance without any feature engineering. Generally, CNNs are hierarchical and RNNs are sequential architectures. The distributed representations, known as word embedding [14], are often used as the first layer in DL models. Distributed representations are mainly learned through context and the learned word vectors can capture general syntactic and semantic information. Those word vectors have proven to be efficient in capturing context and semantic similarity, analogies, and due to its smaller dimensionality are fast and efficient in text mining tasks [15]. Typically, word embeddings are pre-trained by optimizing an auxiliary objective in a large unlabeled corpus and used for other downstream tasks.

A topic model is a probabilistic model, which is used to find the statistics of topics from a large amount of corpus. Traditional topic models such as latent Dirichlet allocation (LDA) [16] and probabilistic latent semantic analysis (PLSA) [17] have been successfully employed in various text corpora. It maps a high dimensional frequency space to a lower-dimensional topic space. Moreover, the topic model can capture semantic information, which can reveal the latent relations among documents. It also can effectively solve the polysemy, synonym, and other problems, which has a vital significance in document feature extraction and content analysis. However, using a topic model to analyze the HoC has not been reported.

Therefore, the purpose of this paper is to focus on the topic analysis for the HoC, design an efficient classification model, capture semantic information from a large amount of documents for extracting complexity of cancer, present a scalable framework to analyze trend and topics for the HoC, and improve the cancer hallmark classification performance and topic modeling result for each hallmark category. The main contributions are as follows.

1. We present a multi-task topic analysis (MTTA) framework to analyze cancer hallmark-specific topics in a multi-task manner by leveraging large amounts of unlabeled documents.
2. We leverage a large number of unlabeled documents related to lung cancer according to the HoC. Experiments on the unlabeled documents have demonstrated that the MTTA framework is potentially valuable for the HoC analysis with an impressive superiority.
3. We highlight the importance of the latent topic models on weak-labeled documents that are produced by the CNN model. The experimental results show that the CNN model can classify cancer hallmarks better than other approaches and conventional topic models can discover topics efficiently.
4. We produce semantic representations for each hallmark category using the CNN model and capture this semantic information for each hallmark category using conventional topic models.

The rest of this paper is organized as follows. Section 2 summarizes related works on cancer hallmarks, topic modeling, and semantic learning. Section 3 presents the MTTA framework and its sub-tasks: cancer hallmark learning (CHL), weak label propagation (WLP), and topic modeling (ToM) in detail. Section 4 outlines the experimental setup and Section 5 shows experimental results of cancer hallmarks classification on manually labeled data and analysis of topic modeling on weak-labeled data related to lung cancer and discusses the effectiveness of the MTTA framework. Finally, we make a conclusion to this paper and discuss the future work in Section 6.

2. Related Work

This section reviews recent advances in biomedical text classification according to the HoC, and topic modeling for cancer research.

2.1. Cancer Hallmarks Classification

Machine learning (ML) and DL techniques have been explosively applied to research in the fields of cancer and NLP. However, few research efforts have addressed cancer hallmark classification with ML and DL.

Baker et al. [18] acquired a collection of PubMed abstracts using a set of search terms representative for each of the 10 hallmarks. They used text mining methodologies such as tokenization, part-of-speech tagging, lemmatization, dependency parsing, and named entity recognition for identifying relevant information in the literature and extracted seven feature types such as a lemmatized bag of words, noun bigrams, grammatical relations, verb classes, named entities, medical subject headings, and chemical lists. They developed 10 independent binary classifiers to predict whether an abstract belongs to each hallmark category and achieved an average F1-score of 77%. Baker et al. [19] presented a joint learning method that learns distributed semantic representations at different types of granularity such as words, sentences, documents, and class using the Support Vector Machine (SVM) algorithm. They also showed a semi-supervised classification result using the same classifier. Their model achieved an F1-score of 76.4% on a corpus of 1580 PubMed abstracts. Baker et al. [20] presented a CNN model to achieve better performance with conventional ML algorithms with manually engineered features. They compared the result with the SVM classifier, outperformed the performance in the seven cancer hallmark tasks. Baker et al. [21] applied a hierarchical model for the multi-label classification task. They initiated the final hidden layer of a neural network that leverages label co-occurrence relations such as hypernymy. They achieved the F1-score of 75.3% on a document level and the accuracy of 89% on the sentence level. Baker et al. [22] developed an automatic text mining methodology and tool to retrieve and organize millions of PubMed literature into the 37 cancer hallmarks by adding subclasses. They used the SVM classifier to predict a biomedical document and trained individual binary classifiers. Their models achieved the average F1-score of 52% and an accuracy of 97.9% on the sentence level classification for all cancer hallmarks.

Du et al. [23] proposed a DL framework for multi-label classification of biomedical text without any manual feature engineering. They trained a single model for a large set of labels. The proposed model alleviates human effort for feature engineering and avoids training an individual model for each class label. Their proposed model achieved the F1-score of 81.5% on a corpus of 1580 abstracts. Pyysalo et al. [24] presented a literature-based discovery system with a particular focus on the molecular biology of cancer using NLP methods and ML. They trained a CNN model on 37 cancer hallmarks including the basic 10 hallmarks. The model supports the system to classify sentences to hallmark categories. Peng et al. [25] evaluated and analyzed biomedical NLP representation models on five tasks with 10 datasets including cancer hallmark classification. They found that the BERT models pre-trained on PubMed abstracts achieved better performance than other models. Their best model achieved the F1-score of 87.3% on a corpus of 1580 PubMed abstracts.

Most of the previous studies used the small dataset, which contains around 1500 documents. In this paper, we leveraged a large number of unlabeled documents related to lung cancer from PubMed.

We analyzed topics on the unlabeled documents for each cancer hallmark category by employing text mining, deep learning, and topic models.

2.2. Topic Modelling for Biomedical Text Mining

Topic modeling is a useful method to enhance cancer researchers' ability to interpret biomedical information. However, using a topic model to analyze the HoC has not been reported, there are few studies that analyzed topics for biomedical and its related documents.

Researchers used a topic model to analyze biomedical literature about genomes such as protein–protein interaction and gene-level analysis. Andrzejewski et al. [26] developed the automatic extraction model to discover the protein–protein interactions from biomedical literature. They employed the LDA model to capture the differences in the vocabulary from Medline abstracts. Wang et al. [27] also employed an LDA generative topic model to extract protein–protein interaction detection from the biomedical literature successfully. They applied the model on a corpus of 5319 full-text documents annotated by MINT and IntAct databases. They reported that the topic model captures the in-depth relationship not only between the methods and related words but also among the different methods. Wang et al. [28] proposed a method to extract commonalities between gene-related documents in an unsupervised manner using a topic model. They employed an LDA model to extract topics from documents. They found that the topics are usually reasonably well described by the currently employed topic algorithms.

Drug discovery is one of the most important challenges in biomedical research. Bisgin et al. [29] investigated the efficacy of topic modeling for the discovery of hidden patterns and their meanings from food and drug administration approved drug labels. They found that the identified topics have distinct contexts directly linked to specific adverse events or therapeutic applications. Bisgin et al. [30] also constructed a probabilistic topic model on the terms in the medical dictionary for drug repositioning. The topic model identified fifty-two unique topics, each containing a set of terms in this study. They found that drugs considered to be similar might often be effective for the same disease.

Retrieving relevant information from biomedical text data is an active challenging area of research. Chen et al. [31] proposed an approach that employs an LDA topic generative model to promoting ranking diversity for biomedical information retrieval. They showed that the proposed approach achieved an 8% improvement over the Aspect MAP reported in TREC 2007 Genomics track [32]. Moreover, Song et al. [33] explored the knowledge structure and trends in bioinformatics by applying text mining techniques including topic modeling to PubMed Central full-text articles. As a result by the topic model, they found that word co-occurrence analysis reveals that major topics focus more on biological aspects than on computational aspects of bioinformatics. Wang et al. [34] developed a literature mining system based on topic modeling called BioTopic. They extracted topics from large-scale documents. They showed that their preprocessing technique improves the result by 5% than traditional preprocessing techniques. They achieved 86% of precision in the topic modeling task. They found that the fine-grained preprocessing with topic modeling shows better results than the previous literature mining systems.

Therefore, there are a few cancers and its related analysis using topic modeling. Cui et al. [35] applied the LDA Gibbs sampling model on the analysis of the top five deadliest cancer research trends, which is extended from the cosine coefficient using the vector space model after transforming the topic-word matrix into a topic word vector. They showed the word clouds to visualize the allocation of the topics for each cancer research, and how to make a computational evaluation for the trend results.

3. Materials and Methods

In this section, we described the presented MTTA framework, which consists of three-main subtasks: CHL, WLP, and TM. The overview of the MTTA framework is shown in Figure 1. Firstly, in the CHL task, we developed the CNN model to learn semantic knowledge about the HoC on a small number of labeled documents. The input of CHL is labeled documents and the output is a trained

model. We used a pre-trained word embedding to capture semantic information as the first layer of the CNN model. In the last layer of the CNN model, we used a softmax function [36] to calculate a probability for each hallmark category. We updated model weights during training. Secondly, in the WLP task, we used the CNN model as our cancer hallmark classifier to annotate unlabeled documents. The input of WLP is unlabeled documents and the output is weak labeled documents. To reduce data noisy, we filtered out the documents by removing a low probability (lower than 80%) after the annotation. As a result, we produced weak-labeled documents. Thirdly, in the ToM task, we employed text mining techniques to preprocess the weak-labeled documents, to identify biomedical entities by discovering medical concepts, and an LDA and LSA models to analyze hallmark-specific topics on the weak-labeled documents. There is no weight updating. The list of abbreviations used in this work are summarized in Table 1.

Table 1. List of abbreviations.

Abbreviations	Definitions
D_L	Labeled documents
N	Number of labeled documents
D_U	Unlabeled documents
S	Number of unlabeled documents
T	Topics
Q	Number of topics
M	Classification model
V	Vocabulary for labeled documents
v	Word vector
D_W	Weak labeled documents
Z	Number of weak labeled documents
P	Class probability for each class (positive and negative)
E	Entities
R	Number of entities
α	Dirichlet prior (Topic distribution)
η	Dirichlet prior (Word distribution)
θ	Multinomial distribution over Q topics
β	Multinomial distribution over R entities

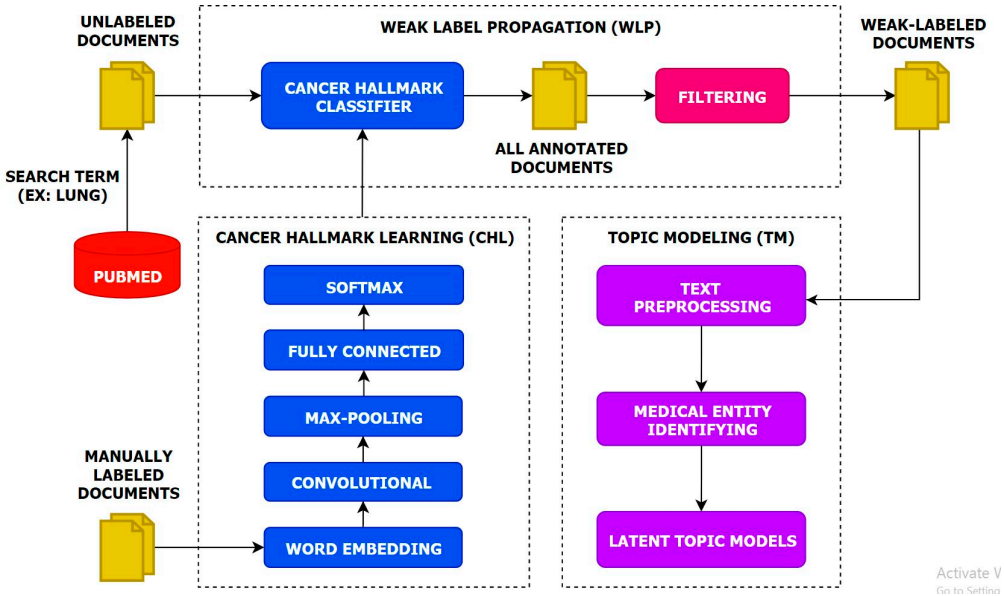

Figure 1. The overview of the MTTA framework.

Algorithm 1 describes the main flow of the MTTA framework as briefly explained above.

Algorithm 1 Multi-task topic analysis framework

Input: A set $D_L = \{D_1, D_2, \ldots, D_N\}$ of N labeled documents; a set $D_U = \{D_1, D_2, \ldots, D_S\}$ of S unlabeled documents;
Output: A set $T = \{T_1, T_2, \ldots, T_Q\}$ of Q topics for each hallmark category;
begin

(1) Train cancer hallmark classifier using CNN model on D_L;
(2) Annotate D_U using pre-trained classifier;
(3) Filter-out annotated D_U using the probability threshold;
(4) Preprocess and identify medical entities from filtered D_U;
(5) Discover T for each hallmark category;

end

3.1. Cancer Hallmarks Learning

In the CHL task, we employed a CNN model to learn the HoC on labeled documents, which has the following five layers: word embedding (200 dim) → convolutional (three kernels with sizes of 3, 4, 5, and 100 feature maps) → max-pooling → fully connected (256 nodes) → softmax. As used in the previous studies [18,20], we used the same strategy to learn the HoC. We trained 10 independent one-class-classifiers for each hallmark category. In the last layer, we used a softmax function to calculate probabilities for each positive and negative class. Positive class indicates specific hallmark category and negative class indicates the non-hallmark category. Algorithm 2 describes the main flow of the CHL task as briefly explained above.

Algorithm 2 Cancer hallmark learning task

Input: A set $D_L = \{D_1, D_2, \ldots, D_N\}$ of N labeled documents; pre-trained word vectors;
Output: A model M trained on D_L;
begin

(1) Create a vocabulary V from D_L;
(2) Retrieve a real-valued vectors v for V;
(3) Initialize the weights for each layer;
(4) Train a classification model;
(5) Calculate probabilities for each positive and negative classes using softmax function;

end

For the cancer hallmarks classification task, those that were rich in biomedical literature were used. Baker et al. [20] investigated the different word embedding techniques for the classification task and compared the performances. As the discussion, the window size 2-word vectors produced by Chiu et al. [37] outperformed all the other vectors. We also applied the same word vector (chiu-win-2) on this task. The chiu-win-2 vectors contain 200-dimensional vectors for 2.2 million words.

3.2. Weak Label Propagation

In this WLP task, we used the pre-trained CNN model as our hallmark classifier, which can classify unlabeled documents into 10 hallmark categories. As calculated by the softmax function, the probability of each class (positive and negative) was produced. To reduce data noise, we filtered out the documents with low probability. We set the threshold value to 0.8. That means the documents with lower and equal probability to 0.8 were ignored and the documents with greater probability than 0.8

were kept for topic analysis. Finally, we produced weak-labeled documents for each hallmark category. Algorithm 3 describes the main flow of the WLP task as briefly explained above.

Algorithm 3 Weak label propagation task

Input: A set $D_U = \{D_1, D_2, \ldots, D_S\}$ of S unlabeled documents; a model M trained on D_L;
Output: A set $D_W = \{D_1, D_2, \ldots, D_Z\}$ of Z weak-labeled documents;
begin

(1) Create a empty list for D_W;
(2) Calculate probabilities P for each class (positive P_{pos} and negative P_{neg});
(3) **if** $P_{pos} > 0.8$ **do**
(4) Append this document to the D_W;
(5) else
(6) Ignore this document;
(7) **end**

end

3.3. Topic Modeling

In the ToM task, we employed conventional topic models to analyze topics for each hallmark category. The ToM task is composed of three steps: (1) text preprocessing; (2) medical entity identification; and (3) latent semantic analysis. Most structured biomedical text data commonly needs classic preprocessing techniques, including data cleaning, stop-word removal, punctuation removal, tokenization, etc. The task of information extraction for medical texts mainly includes named entity recognition. A classic topic modeling has the key steps including the bag of words, model training, and model output. Algorithm 4 describes the main flow of the ToM task as briefly explained above.

Algorithm 4 Topic modeling task

Input: A set $D_W = \{D_1, D_2, \ldots, D_Z\}$ of Z weak-labeled documents;
Output: A set $T = \{T_1, T_2, \ldots, T_Q\}$ of Q topics for each hallmark category;
begin

(1) Preprocess each document in D_W;
(2) Identify biomedical entities form each document;
(3) Apply the conventional topic models to the identified biomedical entities;
(4) Produce a set T for each hallmark category;

end

Each step of the ToM task has been briefly discussed.

3.3.1. Text Preprocessing

For reasons previously mentioned, it is important to clean user-generated data before topic modeling. During preprocessing the following steps were followed for better performance of cancer hallmarks topic modeling.

- All numbers and special characters were removed.
- All uppercase characters were changed into lowercase.
- All non-ASCII character was removed.
- All stopword was removed.

- All classes in an abstract including background, objective, method, result, and conclusion were removed.

3.3.2. Medical Entity Identifier

This section describes the strategy followed by the medical entity identification task. We used the MetaMap tool [38] to identify the unified medical language system (UMLS) [39] terminology in the medical domain and subsequentially, we performed a ranking of relevant documents to be returned. The UMLS is a repository of biomedical vocabularies developed by the US National Library of Medicine. Our work used "2018AA Full Release UMLS Metathesaurus", it contained approximately 3.67 million concepts and 14 million unique concept names from 203 source vocabularies. Vocabularies integrated with the UMLS Metathesaurus include the NCBI taxonomy, Gene Ontology, the Medical Subject Headings (MeSH), MedDRA, RxNorm, or SNOMED-CT. In UMLS when a concept is added to the Metathesaurus, it receives a unique identifier named concept unique identifiers (CUI). This identifier will be very useful in our system. MetaMap is a highly configurable application developed to map biomedical text to the UMLS Metathesaurus. MetaMap employs NLP and computational linguistic techniques: tokenization, part-of-speech tagging, syntactic analysis, word sense disambiguation, and others. This tool first breaks the text into phrases and then, for each phrase, it returns the concepts detected and several other information. Concepts are ranked according to a relevance value. In this paper, we used all the identified CUI, acronyms and abbreviations (AA) as our medical entities. For example, the CUI for the concept "Hypoxia-Inducible Factor" is C0215848. The AA for the concept "Low-folate" is LF. We created a dictionary of the CUI and AA for the following latent topic models.

3.3.3. Latent Topic Models

In this section, we described the conventional latent based topic models LDA and PLSA used in this paper. Basically, it shows the steps in topic modeling, which include a bag of word, training of the model, and output of a topic model. We assumed that there were Q topics, E entities, and Z documents in a corpus.

The bag of words (BoW) model representation is used in information retrieval and natural language processing. In BoW, it represents text by ignoring its order and grammar. As shown in Table 2, there are five entities such as low-folate, metastasis, neoplasm metastasis, decreased folic acid, tryptophanase, and five documents in a corpus. The value E_{ij} in the matrix shows the number of times term i appear in a document. For example, $E_{13} = 8$ means a number of times "low–folate" entity appear in D_3 the document is 8.

Table 2. Bag of words representation of documents.

Entity (CUI or AA)	D_1	D_2	D_3	D_4	D_5
low-folate (LF)	1	2	8	0	7
metastasis (C4255448)	3	4	5	4	9
neoplasm metastasis (C0027627)	3	2	6	3	6
decreased folic acid (C0239623)	9	4	2	1	0
tryptophanase (C0041260)	2	1	0	1	3

The term-document matrix was considered as a simplified representation of documents as the input to the topic model. In a bag of words, the size of the term-document matrix was huge. The goal of topic modeling is to find the concept that runs through documents by analyzing the words of the texts. Topics were discovered during the training of models. For topic models like LDA and PLSA, word term space of the documents converted into topic space as it is definitely sure that topic space is smaller than word term space. The outputs of these models contain two matrices that are the word probability distributions over topics and the second one is the topic probability distributions over documents.

The PLSA is a statistical technique for the analysis of a collection of the document. The probabilistic latent semantic analysis is evolved from a latent semantic analysis. In PLSA, assume D denotes the label of a document i.e., D belongs to $D_W = \{D_1, D_2, \ldots, D_Z\}$, and Z denotes the number of documents in the data. T is a topic, T belongs to $T = \{T_1, T_2, \ldots, T_Q\}$, i.e., there is a Q number of topics. E represents an entity, E belongs to $E = \{E_1, E_2, \ldots, E_R\}$, i.e., there is a R number of entities in data. Therefore, $P(T|D)$ denotes the probability of topic T in document D, and $P(E|T)$ means the probability of entity E in topic T. Then, for PLSA, the generative procedure for each word in the document is described in Algorithm 5.

Algorithm 5 Probabilistic latent semantic analysis

Input: A set $D_W = \{D_1, D_2, \ldots, D_Z\}$ of Z weak-labeled documents; a set $E = \{E_1, E_2, \ldots, E_R\}$ of R entities;
Output: A set $T = \{T_1, T_2, \ldots, T_Q\}$ of Q topics for each hallmark category;
begin

(1) Select a document D_z with probability $P(D_z)$;
(2) Randomly choose a topic T_q from the distribution over topics $P(T_q|D_z)$;
(3) Randomly choose an entity E_r from the distribution over the topic $P(E_r|T_q)$;

end

In LDA, a document is viewed as a combination of different topics where each document is assumed to have a group of topics that are assigned to document using LDA. This is identical to PLSA, in practice, this results in better disambiguation of words and a more precise assignment of documents to topics. LDA is a generalization of the PLSA model, which is equivalent to LDA under a uniform Dirichlet prior to distribution. In LDA, the two probability distributions, $P(T|D)$ and $P(E|T)$ are assumed to be multinomial distributions. Thus, the topic distributions in all documents share the common dirichlet prior α, and the word distributions of topics share the common dirichlet prior η. Given the parameters α and η for document D, parameter θ_D of a multinomial distribution over Q topics is constructed from dirichlet distribution $Dir(\theta_D|\alpha)$. Similarly, for topic Q, parameter β_Q of a multinomial distribution over R entities is derived from Dirichlet distribution $Dir(\beta_Q|\eta)$. As a conjugate prior for the multinomial, the Dirichlet distribution is a convenient choice as a prior and can simplify the statistical inference in LDA. Therefore, in PLSA, by contrast, any common prior probability distribution was not specified for $P(T|D)$ and $P(E|T)$. Naturally, there are no α and η in the generative process of PLSA. Then, for LDA, the procedure for each word in the document is described in Algorithm 6.

Algorithm 6 Latent Dirichlet allocation

Input: A set $D_W = \{D_1, D_2, \ldots, D_Z\}$ of Z weak-labeled documents; a set $E = \{E_1, E_2, \ldots, E_R\}$ of R entities;
Output: A set $T = \{T_1, T_2, \ldots, T_Q\}$ of Q topics for each hallmark category;
begin

(1) Choose θ_D from the distribution $Dir(\alpha)$;
(2) Choose β_Q from the distribution $Dir(\eta)$;
(3) **for each** entity **in** entities **do**
(4) Choose a topic T from the multinomial θ_D;
(5) Choose an entity E from the multinomial β_Q;
(6) **end**

end

4. Experimental Setup

In this section, we described the experimental datasets, hyperparameters of both the DL and topic models and training procedures, respectively. We also described all evaluation measures and baseline methods for each task. Finally, we described the experimental environment all the methods are trained on.

4.1. Dataset

We evaluated the presented MTTA framework on the small manually labeled dataset and crawled a large number of unlabeled documents from PubMed repository.

4.1.1. Manually Labeled Documents

We used the manually labeled cancer hallmark dataset [20], which consists of 1852 biomedical publication abstracts annotated for the hallmarks of cancer by Baker et al. [18]. An annotation was performed at the document level by an expert with more than 15 years of experience in cancer research. We used the training, development and testing data divided in Baker et al. [20]. The statistics of datasets are described in Table 3.

Table 3. Labeled datasets. Y: hallmarks and N: non-hallmarks.

Hallmark	Train		Development		Test	
	Y	**N**	**Y**	**N**	**Y**	**N**
SPS	328	975	43	140	91	275
EGS	172	1131	22	161	46	320
RCD	303	1000	42	141	84	282
ERI	81	1222	11	172	23	343
IA	99	1204	13	170	31	335
AIM	208	1095	29	154	54	312
GIM	227	1076	38	145	68	298
TPI	169	1134	24	159	47	319
DCE	74	1229	10	173	21	345
AID	77	1226	10	173	21	345

4.1.2. Unlabeled Documents

Our interest and a case study's interest are in modeling hallmark-specific topics thus PubMed provided an ideal testing scenario for mining biomedical document abstract data. We used a real-life PubMed dataset related to lung cancer. The data was collected with a search term "lung" via Biopython's module called Entrez [40]. We collected 667,861 unlabeled document abstracts. Manual inspection of the dataset revealed that it was quite compact in terms of topical variance in the hallmark level. Each entry in the dataset represented a single document with its associated abstract including background, objective, methods, results, conclusion, etc.

4.2. Deep Learning Model

This section describes the experimental setup for the deep learning model in the cancer hallmark learning task. For evaluating the CNN model, an exact matching criterion was used to examine three different metrics such as a macro and weighted averaged F1-score. We then compared the CNN model with its supervised manner in terms of the macro and micro F1 score. For this purpose, we implemented the following baseline learning algorithms:

- CNN [12]: a basic CNN model. It has three convolutional layers with a kernel width of 3, 4, and 5 with 100 output channels.
- RCNN [41]: a hybrid RNN and CNN model. The combined architecture CNN followed by Bidirectional Long Short Term Memory (BiLSTM).

- Gated Recurrent Unit (GRU) [42]: one of the basic RNN models. Only forward GRU with 256 hidden units.
- Bidirectional Gated Recurrent Unit (BiGRU) [43]: one of the basic RNN models. Forward and backward GRU with 256 hidden units.
- Long Short Term Memory (LSTM) [44]: one of the basic RNN models. Only forward LSTM with 256 hidden units.
- BiLSTM [45]: one of the basic RNN models. Forward and backward LSTM with 256 hidden units.

The hyperparameters used in the CNN model are described in Table 4. We used Adam optimizer to update parameters while training. We used dropout and an early stopping strategy with patience 20 to avoid overfitting and an early stopping monitored weighted F-score on development sets.

Table 4. Hyperparameters.

Parameter	CNN	RNN
Learning rate	0.001	0.001
Batch size	128	128
Hidden dimension	256	-
Number of layers	2	-
Number of filters	-	100
Filter size	-	[3, 4, 5]
Early stopping patience	20	20
Dropout	0.5	0.5

4.3. Topic Model

For the conventional topic models LDA and PLSA, we used the Gensim Python library [46] with all default parameter settings. The experimental results can be improved by tuning the parameters. We evaluated the models using a semantic coherence score. This metric was proposed to measure the interpretability of topics and was demonstrated to correspond to human coherence judgments. Coherence can also be used for determining the optimal number of topics and the coherence score monotonously decreases if the number of topics increases.

In this paper, the experimental hardware platform was Intel Xeon E3, 32G memory, GTX 1070 Ti. The experimental software platform was Ubuntu 18.04 operating system and the development environment was Python 3.5 programming language. The Pytorch library [47] and the Scikit-learn library [48] of python were used to build the presented MTTA framework and comparative experiments, respectively.

5. Experimental Result and Discussion

In this section, we used the labeled and unlabeled datasets to assess the performance of the MTTA framework, especially, CNN model and topic models, respectively. We mainly compared the performances of methods in two aspects: (1) compare the CNN model with other deep learning-based models on small labeled data in CHC task; (2) annotate the unlabeled documents related to the lung cancer in WLP task; and (3) explore the performance of LDA and PLSA models on the weak-labeled documents in ToM task.

5.1. Cancer Hallmark Classification Result

In the CHC task, we developed the CNN model and the other deep learning-based models. Each model employs the pre-trained chiu-win-2 word vectors using external semantic networks with both base and tuned variants. In the base variant, it uses the word vectors from chiu-win-2 and does not update the word vectors. In the tuned variant, it uses the word vectors and updates word vectors during training for its own specific task. We also compared the models with the best-published performance

in Baker et al. [20] in terms of the macro-averaged F1-score. They used the same CNN model with base and tuned modification of word vectors. As can be seen clearly in Table 5 (macro-averaged F1-score) and Table 6 (micro averaged F1-score).

Table 5. Supervised algorithms on the labeled dataset (macro F1-score).

Model		SPS	EGS	RCD	ERI	IA	AIM	GIM	TPI	DCE	AID
Baker et al. [20]	Best	70.00	71.50	86.90	91.50	85.70	82.60	81.70	84.20	88.30	75.80
CNN	Base	73.49	**82.22**	**92.93**	92.98	88.60	88.45	86.93	87.43	94.36	84.42
	Tuned	74.66	80.15	92.12	**93.18**	**89.10**	**90.15**	**87.62**	87.93	95.83	84.31
RCNN	Base	74.38	81.67	91.10	92.95	86.46	89.82	85.96	**88.92**	**96.62**	**88.61**
	Tuned	75.13	79.77	91.32	92.82	87.41	89.57	86.05	88.70	95.52	88.55
GRU	Base	74.76	79.14	85.95	91.44	88.12	87.78	82.30	86.48	90.48	83.74
	Tuned	**76.65**	74.03	87.86	90.57	88.28	87.38	83.31	86.78	93.52	81.31
BiGRU	Base	70.99	57.45	69.85	87.86	85.46	83.58	80.46	72.99	89.09	74.01
	Tuned	71.16	58.73	70.61	85.87	85.39	82.69	81.05	71.03	90.40	70.37
LSTM	Base	76.47	72.42	82.94	86.32	86.48	86.30	82.22	82.08	91.05	78.83
	Tuned	76.62	69.62	82.84	87.22	87.10	84.84	82.60	82.26	90.43	78.88
BiLSTM	Base	66.23	57.15	58.29	74.86	72.53	68.82	73.99	65.20	81.57	60.82
	Tuned	65.33	56.75	56.36	73.44	72.61	69.73	73.00	64.00	80.90	62.40

Table 6. Supervised algorithms on the labeled dataset (micro F1-score).

Model		SPS	EGS	RCD	ERI	IA	AIM	GIM	TPI	DCE	AID
CNN	Base	81.99	**92.31**	**95.08**	98.13	96.68	94.06	91.68	94.96	98.64	96.92
	Tuned	82.46	91.56	94.61	**98.21**	**96.76**	94.76	**92.15**	95.11	98.91	96.84
RCNN	Base	82.15	91.92	93.94	98.13	95.97	94.96	91.29	**95.31**	**99.10**	97.27
	Tuned	82.46	91.37	94.06	98.09	96.36	**95.00**	91.17	95.27	98.87	**97.35**
GRU	Base	81.98	89.70	90.55	97.99	96.02	93.32	89.67	93.46	97.86	96.70
	Tuned	82.99	89.64	91.68	97.83	95.93	93.13	90.25	93.57	98.52	96.62
BiGRU	Base	78.98	82.55	78.71	97.11	95.25	91.37	88.49	88.27	97.77	94.62
	Tuned	79.51	82.61	78.60	96.84	95.44	90.82	89.18	87.17	97.69	93.98
LSTM	Base	82.97	87.94	88.19	96.95	95.58	92.61	89.48	91.13	98.05	95.88
	Tuned	**83.13**	87.80	88.60	96.92	95.85	91.54	89.78	91.68	97.97	96.13
BiLSTM	Base	76.24	82.52	71.57	94.34	92.78	84.78	85.11	84.97	96.32	92.17
	Tuned	75.63	80.19	70.96	94.10	93.00	85.60	84.95	83.88	96.43	92.64

As shown in the experimental results of the macro-averaged F1 score, the tuned-CNN model outperformed the other models on the four tasks (93.18%, 89.10%, 90.15%, and 87.62% on ERI, IA, AIM, and GIM, respectively). The base-CNN model outperformed the other models on the two tasks (82.22% and 92.93% on EGS and RCD, respectively). The combined model base-RCNN outperformed on the three tasks (88.92%, 96.62%, and 88.61% on TPI, DCE, and AID, respectively). The tuned-GRU model achieved 76.65% on the remained SPS task.

As shown in the experimental results of the micro-averaged F1 score, the tuned-CNN model outperformed the other models on the three tasks (98.21%, 96.76%, and 92.15% on ERI, IA, and GIM, respectively). The base-CNN model outperformed the other models on the two tasks (92.31% and 95.08% on EGS and RCD, respectively). The combined model base-RCNN outperformed on the two tasks (95.31% and 99.10% on TPI and DCE, respectively). The combined model tuned-RCNN outperformed on the remaining two tasks (95.00% and 97.35% on AIM and AID, respectively). The tuned-LSTM model achieved 83.13% on the remaining SPS task.

We concluded that for this experiment, CNN based models were more efficient than RNN based models in case of a low semantic dependency and a high out of vocabulary tasks. The CNN models were good at extracting local and position invariant features. For the weak label propagation task, we selected the tuned-CNN model because of its performance.

5.2. Weak Label Propagation Result

In this task, we show the experimental results of weak-labeled documents related to lung cancer. We annotated a large number of unlabeled documents using the tuned-CNN model showed in the previous section. While label propagation, the softmax function was used to calculate a probability for each positive and negative class. To reduce the data noise problem, we filtered-out the documents that had low probability (lower than 0.8) and selected the documents that had a high probability (higher and equal to 0.8). The statistics of the weak-labeled data are shown in Table 7.

We found 48,953 documents related to ERI cancer hallmark from a total of 667,861 documents at most. It had a total of 7,698,235 medical concepts (CUI) and 80,444 unique CUIs. We found only 43 documents related to DCE cancer hallmark from a total of 667,861 documents at least. It has a total of 7701 CUIs and 2124 unique CUIs.

After finding all hallmark specific documents, we applied the conventional topic models LDA and PLSA on the weak-labeled documents.

Table 7. The statistics of the weak-labeled documents (lung cancer).

Hallmark	No. of Documents	No. of Entities	No. of Unique Entities
SPS	4097	766,146	22,576
EGS	950	170,965	11,272
RCD	14,313	2,710,563	56,052
ERI	604	102,357	9531
IA	3009	582,378	22,362
AIM	48,953	7,698,235	80,444
GIM	7733	1,444,839	33,579
TPI	30,403	6,006,649	67,733
DCE	43	7701	2124
AID	13,098	2,094,029	40,055

5.3. Topic Modelling Result

This section shows the results of the topic modeling. We reported the top-10 concepts from the top-1 topic for each hallmark in Tables 8–17. The topics of each hallmark were explored using LDA and PLSA models and the results were compared. We found the following topics were most related to cancer hallmarks from lung cancer data. The common non-medical words are highlighted.

As shown in Table 8, the identified CUI "Epidermal Growth Factor Receptor" is mostly concerned with the hallmark "sustaining proliferative signaling" (SPS) in the different concepts. For example, EGFR is a shortage of the epidermal growth factor receptor. The concepts "*EGFR* gene, EGFR protein, and EGFR measurement" were found in the top-10 concepts. Liu et al. [49] reported that activation of EGFR-tyrosine kinases is a key reason for lung cancer progression. We found that EGFR was the best topic in the SPS hallmark.

As shown in Table 9, *TP53* was mostly concerned about the hallmark evading growth suppressors (EGSs) in multiple concepts such as the *TP53* wt Allele and *TP53* gene. Amin et al. [50] mentioned the recent report of the Cancer Genome Atlas (TCGA) assessment of squamous cell lung cancers, where the most common significantly mutated gene was *TP53*. We found that *TP53* was the best topic in the EGS hallmark.

As shown in Table 10, we found that the concept "Apoptosis" was the key topic in resisting cell death (RCD). Liu et al. [51] discussed the role of apoptosis in non-small-cell lung cancer (NSCLC). They report that the processes of autophagy and apoptosis, which induce degradation of proteins and organelles or cell death upon cellular stress.

As shown in Table 11, we found that the concept "senility", "senescence", and "old age" were the key concepts in the hallmark enabling replicative immortality (ERI). Senescence or biological aging is the gradual deterioration of functional characteristics and senility describes that a person who is experiencing dementia brought about by old age. Yaswen et al. [52] discussed the therapeutic targeting

of replicative immortality. They reported that a protective role of senescence has been inferred in murine models of lung adenomas, T-cell lymphomas, prostate tumors, and pituitary tumors.

As shown in Table 12, we found that the concept "Vascular Endothelial Growth Factors (VEGF)" was a key concept in the hallmark inducing angiogenesis (IA). Shimoyamada et al. [53] reported that VEGF is crucial for angiogenesis, vascular permeability, and metastasis during tumor development.

As shown in Table 13, we found that the concepts "Neoplasm" and "Metastasis" were mostly concerned about the hallmark activating invasion and metastasis (AIM). Martin et al. [54] reported that the liver is one of the most common sites for metastatic disease and in the United States and Europe, secondary liver neoplasms are far more common than primary hepatic neoplasms. Lung cancer that spreads to the liver is called metastatic lung cancer rather than liver cancer.

As shown in Table 14, we found that the concepts "mutation" and "mutation abnormality" were the key topics related to the hallmark genome instability and mutation (GIM). Ninomiya et al. [55] discussed genetic instability in lung cancer. Genetic instability refers to a high frequency of mutations within the genome of a cellular lineage. These mutations can include changes in nucleic acid sequences, chromosomal rearrangements, or aneuploidy.

As shown in Table 15, we found that the concept "lipopolysaccharides" was the key concept related to the hallmark tumor-promoting inflammation (TPI). Melkamu et al. [56] discussed that lipopolysaccharide enhances mouse lung tumorigenesis. Lipopolysaccharide (LPS), known as a trigger of inflammatory responses, has been suggested to be implicated in cancer invasion or angiogenesis [57].

As shown in Table 16, we found that the concept of "aerobic glycolysis" was the key concept related to the hallmark deregulating cellular energetics (DCE). Min et al. [58] discussed metabolic alterations in NSCLC and the impact of metabolic reprogramming on the development and progression of human cancers and deregulated metabolism. They mentioned that aerobic glycolysis is important for reducing the economic and social burden of cancer.

As shown in Table 17, we found that the concept "Blood group antibody I" was the key concept related to the hallmark avoiding immune destruction (AID). Gwin et al. [59] discussed the loss of blood group antigen A in NSCLC. They confirmed the finding, NSCLC patients who were blood group A and had paraffin-embedded primary lung cancer tissue suitable for immunohistological analysis of antigen A expression.

Table 8. The top-1 topic on the sustaining proliferative signaling (SPS) hallmark.

LDA	Patients; Epidermal Growth Factor Receptor Measurement; Epidermal Growth Factor Receptor; EGFR protein, human; Combined; Therapeutic procedure; Tryptophanase; Non-Small Cell Lung Carcinoma; EGFR gene; combination of objects.
PLSA	Epidermal Growth Factor Receptor Measurement; EGFR protein, human; Epidermal Growth Factor Receptor; EGFR gene; Non-Small Cell Lung Carcinoma; Patients; Tryptophanase; Therapeutic procedure; Non-Small Cell Lung Cancer Pathway; NCI CTEP SDC Non-Small Cell Lung Cancer Sub-Category Terminology.

Table 9. The top-1 topic on the evading growth suppressor (EGS) hallmark.

LDA	Induce (action); Expression (foundation metadata concept); Expression procedure; Effect; Homo sapiens; Apoptosis; Tryptophanase; *TP53* wt Allele; *TP53* gene; Inhibition.
PLSA	Cell Count; Induce (action); Expression procedure; Expression (foundation metadata concept); Tryptophanase; Carcinoma of lung; Apoptosis; Malignant neoplasm of lung; Effect; *TP53* gene.

Table 10. The top-1 topic on the resisting cell death (RCD) hallmark.

LDA	Tryptophanase; Neoplasms; Treating; Therapeutic procedure; Patients; PSA Level Less than Two; 2+ Score, WHO; 2+ Score; therapeutic aspects; Administration procedure.
PLSA	Apoptosis; Induce (action); Tryptophanase; Cell Count; Expression procedure; Effect; Expression (foundation metadata concept); Therapeutic procedure; Treating; Increase.

Table 11. The top-1 topic on the enabling replicative immortality (ERI) hallmark.

LDA	Senility; Old age; Induce (action); Cellular Senescence; Tryptophanase; Fibroblasts; Cell Count; Homo sapiens; Associated with; Increase.
PLSA	Old age; Senility; Induce (action); Cellular Senescence; Tryptophanase; Cell Count; Expression procedure; Expression (foundation metadata concept); Fibroblasts; Homo sapiens.

Table 12. The top-1 topic on the inducing angiogenesis (IA) hallmark.

LDA.	Vascular Endothelial Growth Factors; Recombinant Vascular Endothelial Growth Factor; Neoplasms; Tryptophanase; Angiogenic Process; Laboratory mice; Social group; Group Object; Tumor Mass; Population Group.
PLSA	Vascular Endothelial Growth Factors; Recombinant Vascular Endothelial Growth Factor; Tumor Angiogenesis; Angiogenic Process; Tryptophanase; Expression procedure; Expression (foundation metadata concept); Neoplasms; Patients; P prime.

Table 13. The top-1 topic on the activating invasion and metastasis (AIM) hallmark.

LDA	Cell Count; Neoplasm Metastasis; Secondary Neoplasm; Tryptophanase; Metastatic to; Metastatic Neoplasm; metastatic qualifier; Metastatic Disease Clinical Trial Setting; Neoplasms; Metastasis.
PLSA	Patients; Neoplasm Metastasis; Secondary Neoplasm; Tryptophanase; Metastatic Neoplasm; Neoplasms; Metastasis; P Blood group antibodies; P prime; Expression procedure.

Table 14. The top-1 topic on the genome instability and mutation (GIM) hallmark.

LDA	Mutation; Present; Tryptophanase; adduct; 1+ Score, WHO; 1+ Score; Greater than one; Carcinoma of lung; Mutation Abnormality; Malignant neoplasm of lung.
PLSA	Mutation; Patients; Tryptophanase; Mutation Abnormality; P Blood group antibodies; P prime; Induce (action); Exposure to; Present; EGFR protein, human.

Table 15. The top-1 topic on the tumor-promoting inflammation (TPI) hallmark.

LDA	Lipopolysaccharides; Tumor Necrosis Factor-alpha; Induce (action); TNF protein, human; Increase; Alpha tumor necrosis factor measurement; cytokine; Interleukin-1 beta; Protons; Hepatic Involvement.
PLSA	Population Group; Groups; Social group; User Group; Stage Grouping; Group Object; Tryptophanase; Induce (action); Increase; House mice.

Table 16. The top-1 topic on the deregulating cellular energetics (DCE) hallmark.

LDA	Aerobic glycolysis; Inhibition; Tryptophanase; Glycolysis; Induce (action); Metabolic Process, Cellular; Glucose; Expression (foundation metadata concept); Increase; Mitochondria.
PLSA	Aerobic glycolysis; Glycolysis; Tryptophanase; Expression procedure; Expression (foundation metadata concept); Cell Count; Increase; production; Malignant Neoplasms; Induce (action).

Table 17. The top-1 topic on the avoiding immune destruction (AID) hallmark.

LDA	Blood group antibody I; Iodides; Tryptophanase; Neoplasms; Tumor Mass; Specimen Source Codes—tumor; Neoplasm Metastasis; Induce (action); Vaccination; T-Lymphocyte.
PLSA	Tryptophanase; Patients; House mice; Laboratory mice; T-Lymphocyte; SNAP25 wt Allele; SNAP25 protein, human; *HERPUD1* gene; HERPUD1 wt Allele; Negation.

5.4. Visualization

This section reports the visual representation of the top-1 topic for each hallmark as a word cloud. Word clouds have emerged as a straightforward and visually appealing visualization method for text. They are used in various contexts as a means to provide an overview by distilling text down to those words that appear with the highest frequency. Typically, this is done in a static way as pure text summarization.

We created the word clouds on the top-100 concepts for each hallmark category. We used the "WordCloud" python library to generate all word clouds. The word clouds for each hallmark category are shown in Figures 2–11. The left side of each figure is a result of the LDA model and the right side of each figure is a result of the PLSA model. See Tables 8–17 related to Figures 2–11.

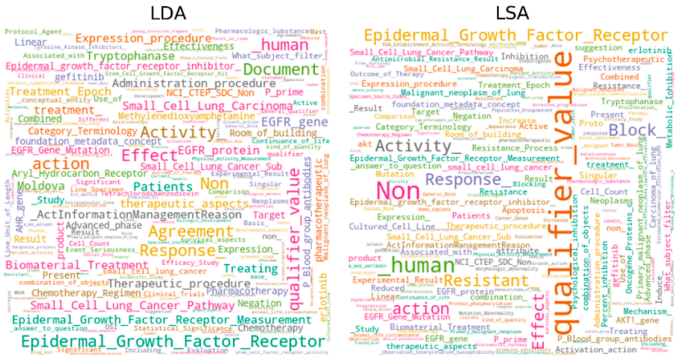

Figure 2. Word cloud of the top-1 topics on the SPS hallmark.

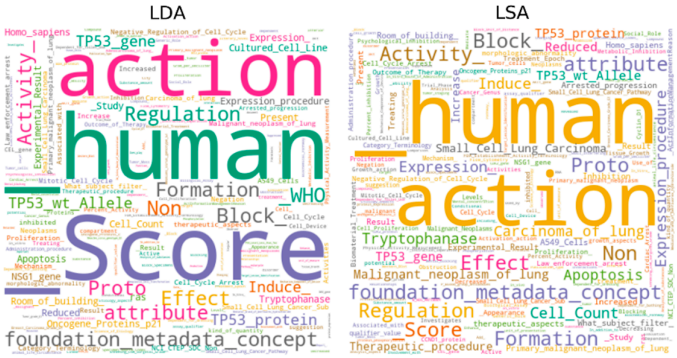

Figure 3. Word cloud of the top-1 topics on the EGS hallmark.

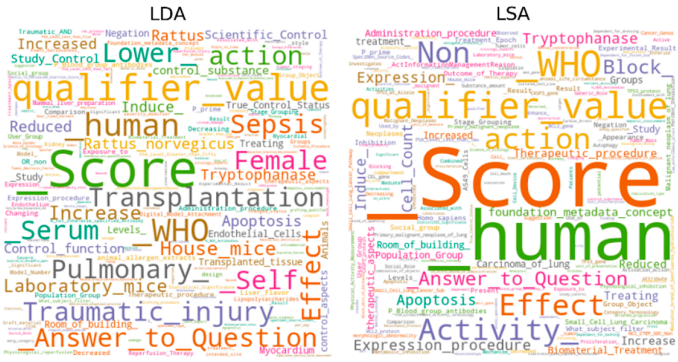

Figure 4. Word cloud of the top-1 topics on the RCD hallmark.

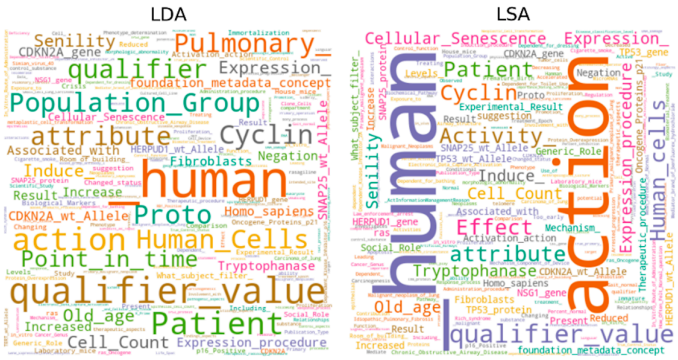

Figure 5. Word cloud of the top-1 topics on the ERI hallmark.

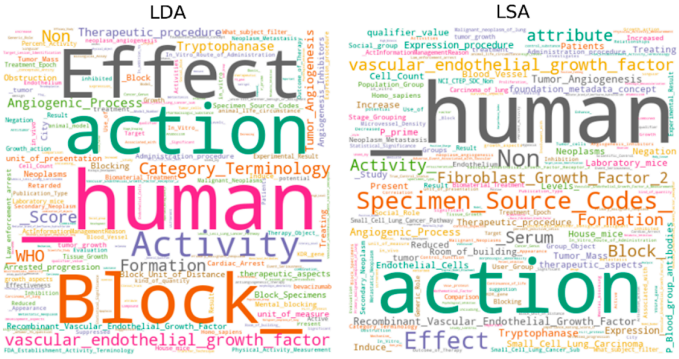

Figure 6. Word cloud of the top-1 topics on the IA hallmark.

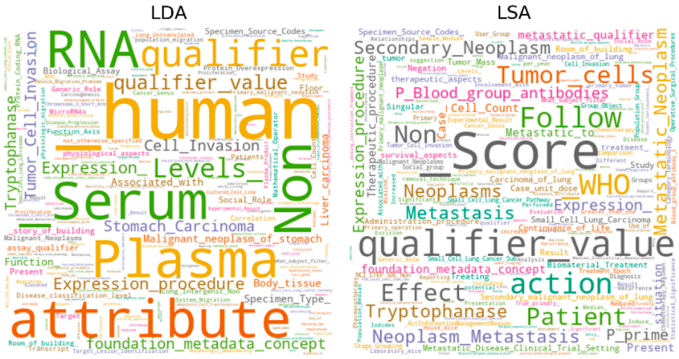

Figure 7. Word cloud of the top-1 topics on the AIM hallmark.

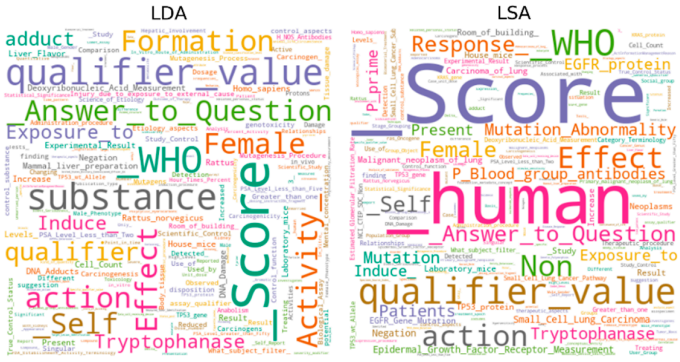

Figure 8. Word cloud of the top-1 topics on the GIM hallmark.

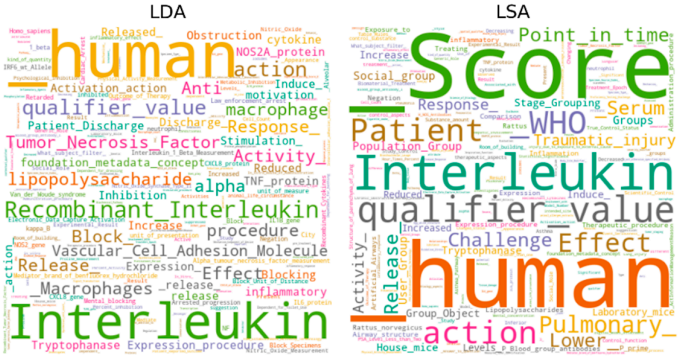

Figure 9. Word cloud of the top-1 topics on the TPI hallmark.

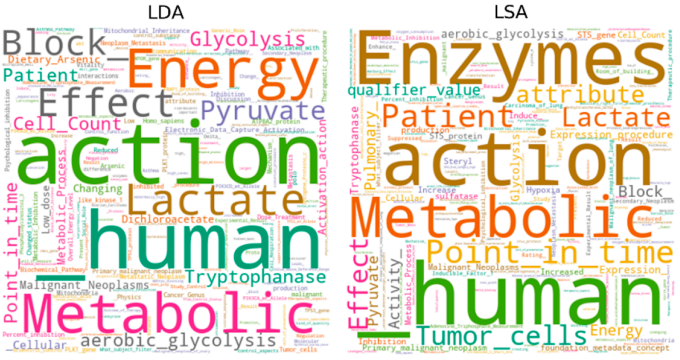

Figure 10. Word cloud of the top-1 topics on the DCE hallmark.

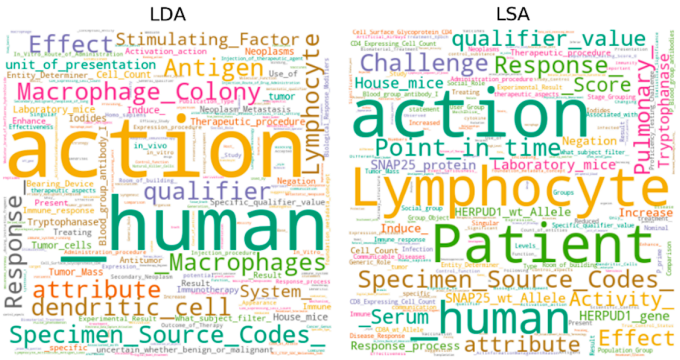

Figure 11. Word cloud of the top-1 topics on the AID hallmark.

5.5. Optimal Number of Topics

This section explores the optimal number of topics in terms of the coherence score. We trained the LDA and PLSA models by setting the number of topics to 20. As can be seen clearly in Figures 12–21, our trained models were compared by coherence scores on each number of topics.

As shown in Figure 12, for hallmark SPS, the optimal number of topics was 5 explored by the PLSA model. It achieved 51.70 of the coherence score. As shown in Figure 13, for hallmark EGS, the optimal number of topics was 10 explored by the PLSA model. It achieved 44.61 of the coherence score. As shown in Figure 14, for hallmark RCD, the optimal number of topics was 19 explored by the LDA model. It achieved 49.44 of the coherence score. As shown in Figure 15, for hallmark ERI, the optimal number of topics was 9 explored by the PLSA model. It achieved 37.90 of the coherence score. As shown in Figure 16, for hallmark IA, the optimal number of topics was 5 explored by the PLSA model. It achieved 47.39 of the coherence score. As shown in Figure 17, for hallmark AIM, the optimal number of topics was 13 explored by the LDA model. It achieved 52.95 of the coherence score. As shown in Figure 18, for hallmark GIM, the optimal number of topics was 5 explored by the LDA model. It achieved 50.00 of coherence score. As shown in Figure 19, for hallmark TPI, the optimal number of topics was 17 explored by the LDA model. It achieved 49.44 of the coherence score. As shown in Figure 20, for hallmark DCE, the optimal number of topics was 6 explored by the PLSA model. It achieved 45.75 of the coherence score. As shown in Figure 21, for hallmark AID, the optimal number of topics was 17 explored by the LDA model. It achieved 52.18 of the coherence score.

As a result, the LDA model performed better results than the PLSA model on five hallmark tasks (RCD, AIM, GIM, TPI, and AID) in terms of the coherence score. The PLSA model performed better results than the LDA model on five hallmark tasks (SPS, EGS, ERI, IA, and DCE) in terms of the coherence score.

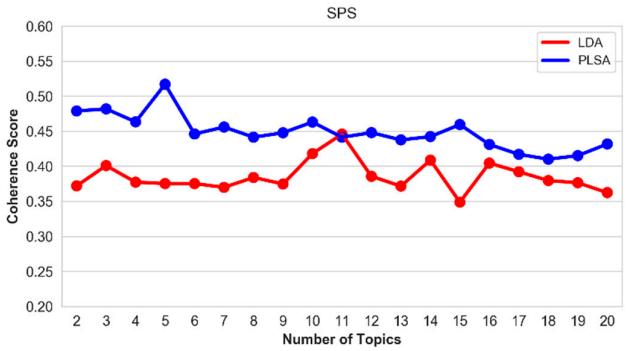

Figure 12. Coherence score of the top-20 topics on the SPS hallmark.

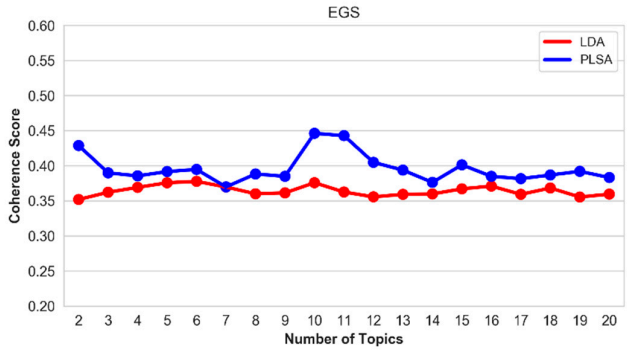

Figure 13. Coherence score of the top-20 topics on the EGS hallmark.

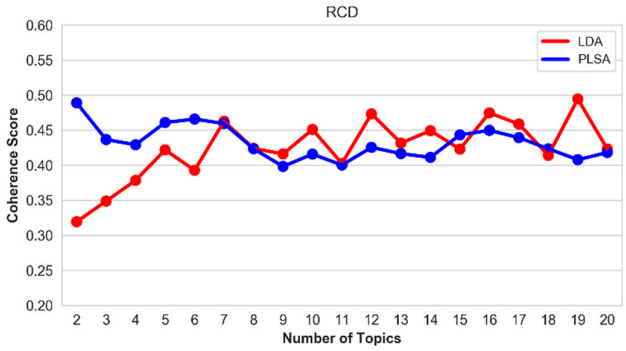

Figure 14. Coherence score of the top-20 topics on the RCD hallmark.

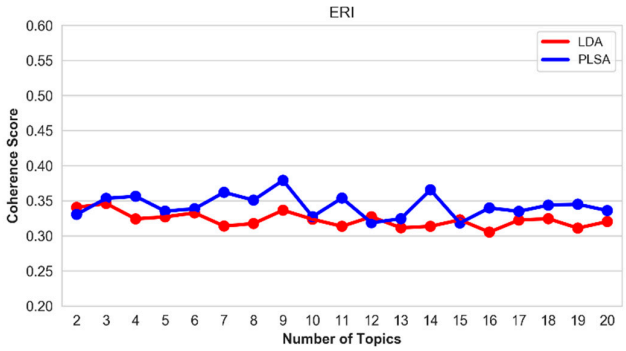

Figure 15. Coherence score of the top-20 topics on the ERI hallmark.

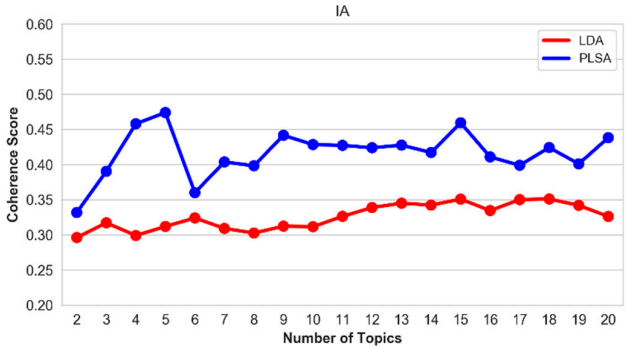

Figure 16. Coherence score of the top-20 topics on the IA hallmark.

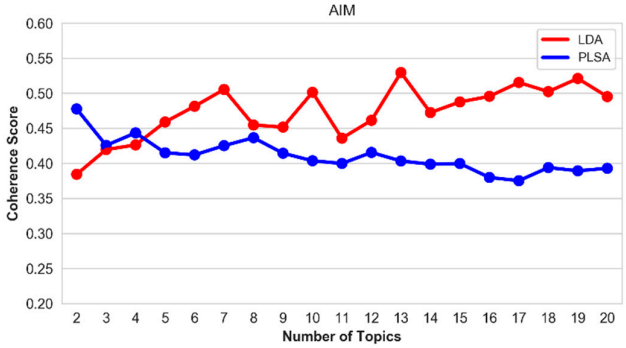

Figure 17. Coherence score of the top-20 topics on the AIM hallmark.

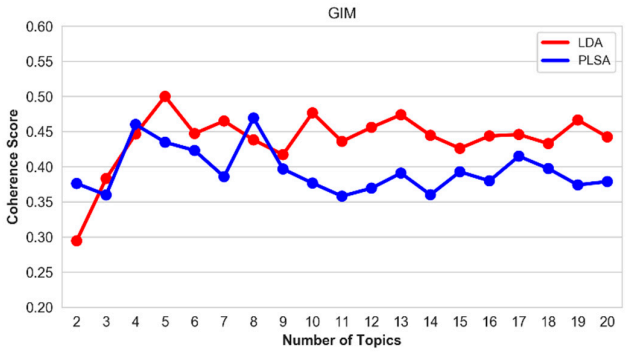

Figure 18. Coherence score of the top-20 topics on the GIM hallmark.

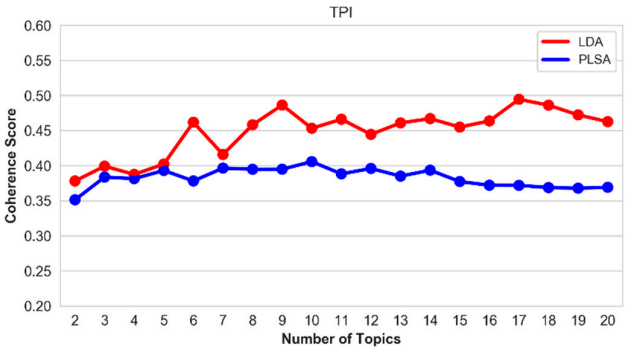

Figure 19. Coherence score of the top-20 topics on the TPI hallmark.

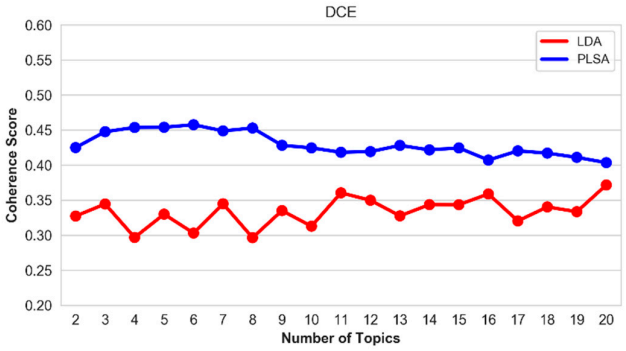

Figure 20. Coherence score of the top-20 topics on the DCE hallmark.

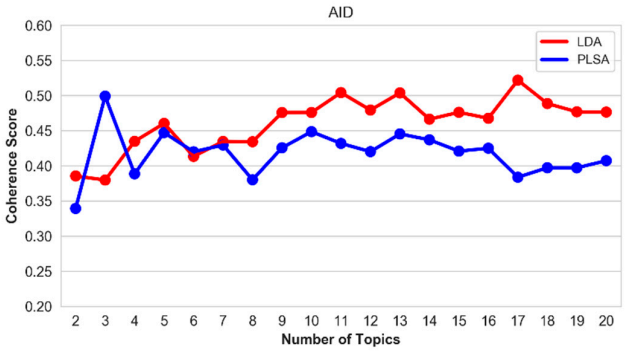

Figure 21. Coherence score of the top-20 topics on the AID hallmark.

6. Conclusions and Future Work

This paper presented a topic analysis framework, called MTTA in a multi-task manner. The MTTA framework consists of three main tasks called CHL, WLP, and ToM. The CHL task employed the

deep learning-based supervised learning algorithm, which could learn cancer hallmarks on the existing manually labeled dataset. We compared multiple deep learning models and then selected the tuned-CNN model in this paper. The WLP task employed the pre-trained CNN model as a cancer hallmark classifier for unlabeled documents. To reduce data noise, we used a simple threshold on each class probability calculated by softmax function. Finally, we created weak-labeled documents for topic analysis. The ToM task employed conventional topic models LDA and PLSA for comparison. The topic models utilize only biomedical concepts identified by UMLS terminology. The presented MTTA framework was highly scalable on a large number of unlabeled documents. We found that the conventional topic models were efficient to analyze topics on the weak-labeled document according to the HoC. We verified that the CNN models achieved better results than other deep learning models.

We studied a large number of documents related to lung cancer for topic analysis. The pre-trained deep learning model produced hallmark specific weak-labeled documents and topic models discovered hallmark specific topics. The results show that we could efficiently extract a complex structure of cancer knowledge according to the cancer hallmark and its related topics.

For future work, we will avoid using additional feature engineering and text mining methods for improving the performance and usability of this work in an end-to-end manner. We will improve the performance of cancer hallmark learning using the transformer networks, for example, the BERT [25] pre-trained model. It will improve the quality of weak-labeled documents. To avoid class imbalance problems that introduce bias and developing too many models, we will develop a single model, which addresses multi-label classification task. We would point out that the conventional topic models LDA and PLSA did not fit the medical concepts very well so we planned to investigate other topic models. We will also compare different pre-processing techniques to identify medical entities from a large number of unlabeled documents.

Author Contributions: Conceptualization, E.B. and K.H.R.; Methodology, E.B.; Formal analysis, E.B., V.-H.P. and K.H.R.; Investigation, E.B.; Data Curation, E.B.; Writing—original draft preparation, E.B.; Writing—review and editing, V.-H.P. and K.H.R.; Supervision, K.H.R.; Project administration, K.H.R.; Funding acquisition, K.H.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT &, Future Planning (No.2017R1A2B4010826 and NRF-2019K2A9A2A06020672).

Acknowledgments: The authors would like to thank reviewers for their suggestions to improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mehmet Sitki Copur, M.D. State of Cancer Research around the Globe. *Oncology* **2019**, *14*, 33.
2. Hanahan, D.; Weinberg, R.A. The hallmarks of cancer. *Cell* **2000**, *100*, 57–70. [CrossRef]
3. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [CrossRef]
4. Gutschner, T.; Diederichs, S. The hallmarks of cancer: A long non-coding RNA point of view. *RNA Biol.* **2012**, *9*, 703–719. [CrossRef] [PubMed]
5. Piao, Y.; Piao, M.; Ryu, K.H. Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles. *Comput. Biol. Med.* **2017**, *80*, 39–44. [CrossRef] [PubMed]
6. Li, F.; Piao, M.; Piao, Y.; Li, M.; Ryu, K.H. A New direction of cancer classification: Positive effect of Low-ranking MicroRNAs. *Osong Public Health Res. Perspect.* **2014**, *5*, 279–285. [CrossRef] [PubMed]
7. Munkhdalai, T.; Li, M.; Batsuren, K.; Park, H.A.; Choi, N.H.; Ryu, K.H. Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations. *J. Chemin.* **2015**, *7*, 9. [CrossRef]
8. Munkhdalai, T.; Namsrai, O.E.; Ryu, K.H. Self-training in significance space of support vectors for imbalanced biomedical event data. *BMC Bioinform.* **2015**, *16*, 6. [CrossRef]
9. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]

10. He, L.; Lee, K.; Lewis, M.; Zettlemoyer, L. Deep semantic role labeling: What works and what's next. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 473–483.
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; 2017; pp. 5998–6008.
12. Kim, Y. Convolutional neural networks for sentence classification. *arXiv* **2014**, arXiv:1408.5882.
13. Mikolov, T.; Karafiát, M.; Burget, L.; Černocký, J.; Khudanpur, S. Recurrent neural network based language model. In Proceedings of the Eleventh Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010.
14. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2013; pp. 3111–3119.
15. Batbaatar, E.; Li, M.; Ryu, K.H. Semantic-emotion neural network for emotion recognition from text. *IEEE Access* **2019**, *7*, 111866–111878. [CrossRef]
16. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
17. Hofmann, T. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.* **2001**, *42*, 177–196. [CrossRef]
18. Baker, S.; Silins, I.; Guo, Y.; Ali, I.; Högberg, J.; Stenius, U.; Korhonen, A. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics* **2015**, *32*, 432–440. [CrossRef] [PubMed]
19. Baker, S.; Kiela, D.; Korhonen, A. Robust text classification for sparsely labelled data using multi-level embeddings. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 2333–2343.
20. Baker, S.; Korhonen, A.; Pyysalo, S. Cancer hallmark text classification using convolutional neural networks. In Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), Osaka, Japan, 11–16 December 2016; pp. 1–9.
21. Baker, S.; Korhonen, A. *Initializing Neural Networks for Hierarchical Multi-Label Text Classification*; BioNLP: Vancouver, BC, Canada, 2017; pp. 307–315.
22. Baker, S.; Ali, I.; Silins, I.; Pyysalo, S.; Guo, Y.; Högberg, J.; Stenius, U.; Korhonen, A. Cancer Hallmarks Analytics Tool (CHAT): A text mining approach to organize and evaluate scientific literature on cancer. *Bioinformatics* **2017**, *33*, 3973–3981. [CrossRef]
23. Du, J.; Chen, Q.; Peng, Y.; Xiang, Y.; Tao, C.; Lu, Z. ML-Net: Multi-label classification of biomedical texts with deep neural networks. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1279–1285. [CrossRef]
24. Pyysalo, S.; Baker, S.; Ali, I.; Haselwimmer, S.; Shah, T.; Young, A.; Guo, Y.; Högberg, J.; Stenius, U.; Narita, M. LION LBD: A literature-based discovery system for cancer biology. *Bioinformatics* **2018**, *35*, 1553–1561. [CrossRef]
25. Peng, Y.; Yan, S.; Lu, Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *arXiv* **2019**, arXiv:1906.05474.
26. Andrzejewski, D. *Modeling Protein–Protein Interactions in Biomedical Abstracts with Latent Dirichlet Allocation*; CS 838-Final Project; University of Wisconsin–Madison: Madison, WI, USA, 2006.
27. Wang, H.; Huang, M.; Zhu, X. Extract interaction detection methods from the biological literature. *BMC Bioinform.* **2009**, *10*, 55. [CrossRef]
28. Wang, V.; Xi, L.; Enayetallah, A.; Fauman, E.; Ziemek, D. GeneTopics-interpretation of gene sets via literature-driven topic models. *BMC Syst. Biol.* **2013**, *7*, 10. [CrossRef]
29. Bisgin, H.; Liu, Z.; Fang, H.; Xu, X.; Tong, W. Mining FDA drug labels using an unsupervised learning technique-topic modeling. *BMC Bioinform.* **2011**, *12*, 11. [CrossRef] [PubMed]
30. Bisgin, H.; Liu, Z.; Kelly, R.; Fang, H.; Xu, X.; Tong, W. Investigating drug repositioning opportunities in FDA drug labels through topic modeling. *BMC Bioinform.* **2012**, *13*, 6. [CrossRef] [PubMed]
31. Chen, Y.; Yin, X.; Li, Z.; Hu, X.; Huang, J.X. A LDA-based approach to promoting ranking diversity for genomics information retrieval. *BMC Genomics* **2012**, *13*, 2. [CrossRef] [PubMed]
32. Hersh, W.R.; Cohen, A.M.; Roberts, P.M.; Rekapalli, H.K. *TREC 2006 Genomics Track Overview*; TREC: Gaithersburg, MD, USA, 2006.

33. Song, M.; Kim, S.Y. Detecting the knowledge structure of bioinformatics by mining full-text collections. *Scientometrics* **2013**, *96*, 183–201. [CrossRef]
34. Wang, X.; Zhu, P.; Liu, T.; Xu, K. BioTopic: A topic-driven biological literature mining system. *Int. J. Data Min. Bioinform.* **2016**, *14*, 373–386. [CrossRef]
35. Cui, M.; Liang, Y.; Li, Y.; Guan, R. Exploring Trends of Cancer Research Based on Topic Model. *IWOST-1* **2015**, *1339*, 7–18.
36. Dunne, R.A.; Campbell, N.A. On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function. In Proceedings of the 8th Australian Conference on Neural Networks, Canberra, Australia, 10–12 April 1997; Volume 181, p. 185.
37. Chiu, B.; Crichton, G.; Korhonen, A.; Pyysalo, S. How to train good word embeddings for biomedical NLP. In Proceedings of the 15th Workshop on Biomedical Natural Language Processing, Berlin, Germany, 12 August 2016; pp. 166–174.
38. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In Proceedings of the AMIA Symposium. American Medical Informatics Association, Chicago, IL, USA, 14–18 November 2001; p. 17.
39. Bodenreider, O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, 267–270. [CrossRef]
40. Chapman, B.; Chang, J. Biopython: Python tools for computational biology. *ACM Sigbio Newsl.* **2000**, *20*, 15–19. [CrossRef]
41. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. In Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
42. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
43. Luo, X.; Zhou, W.; Wang, W.; Zhu, Y.; Deng, J. Attention-based relation extraction with bidirectional gated recurrent unit and highway network in the analysis of geological data. *IEEE Access* **2017**, *6*, 5705–5715. [CrossRef]
44. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
45. Liwicki, M.; Graves, A.; Fernàndez, S.; Bunke, H.; Schmidhuber, J. A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007, Curitiba, Brazil, 23–26 September 2007.
46. Řehůřek, R.; Sojka, P. *Gensim—Statistical Semantics in Python. Statistical Semantics; Gensim*; EuroScipy: Paris, France, 2011.
47. Ketkar, N. *Introduction to Pytorch*; Apress: Berkeley, CA, USA, 2017.
48. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
49. Liu, T.C.; Jin, X.; Wang, Y.; Wang, K. Role of epidermal growth factor receptor in lung cancer and targeted therapies. *Am. J. Cancer Res.* **2017**, *7*, 187. [PubMed]
50. Amin, A.R.; Karpowicz, P.A.; Carey, T.E.; Arbiser, J.; Nahta, R.; Chen, Z.G.; Dong, J.T.; Kucuk, O.; Khan, G.N.; Huang, G.S. Evasion of anti-growth signaling: A key step in tumorigenesis and potential target for treatment and prophylaxis by natural compounds. In *Seminars in Cancer Biology*; Elsevier: Amsterdam, The Netherlands, 2015; Volume 35, pp. 55–77.
51. Liu, G.; Pei, F.; Yang, F.; Li, L.; Amin, A.D.; Liu, S.; Buchan, J.R.; Cho, W.C. Role of autophagy and apoptosis in non-small-cell lung cancer. *Int. J. Mol. Sci.* **2017**, *18*, 367. [CrossRef]
52. Yaswen, P.; MacKenzie, K.L.; Keith, W.N.; Hentosh, P.; Rodier, F.; Zhu, J.; Firestone, G.L.; Matheu, A.; Carnero, A.; Bilsland, A. Therapeutic targeting of replicative immortality. In *Seminars in Cancer Biology*; Elsevier: Amsterdam, The Netherlands, 2015; Volume 35, pp. 104–128.
53. Shimoyamada, H.; Yazawa, T.; Sato, H.; Okudela, K.; Ishii, J.; Sakaeda, M.; Kashiwagi, K.; Suzuki, T.; Mitsui, H.; Woo, T. Early growth response-1 induces and enhances vascular endothelial growth factor-A expression in lung cancer cells. *Am. J. Pathol.* **2010**, *177*, 70–83. [CrossRef]
54. Martin, T.A.; Ye, L.; Sanders, A.J.; Lane, J.; Jiang, W.G. Cancer Invasion and Metastasis: Molecular and Cellular Perspective. Available online: https://www.ncbi.nlm.nih.gov/books/NBK164700/(accessedonine30December2019) (accessed on 30 December 2019).

55. Ninomiya, H.; Nomura, K.; Satoh, Y.; Okumura, S.; Nakagawa, K.; Fujiwara, M.; Tsuchiya, E.; Ishikawa, Y. Genetic instability in lung cancer: Concurrent analysis of chromosomal, mini-and microsatellite instability and loss of heterozygosity. *Br. J. Cancer* **2006**, *94*, 1485. [CrossRef]
56. Melkamu, T.; Qian, X.; Upadhyaya, P.; O'Sullivan, M.G.; Kassie, F. Lipopolysaccharide enhances mouse lung tumorigenesis: A model for inflammation-driven lung cancer. *Vet. Pathol.* **2013**, *50*, 895–902. [CrossRef]
57. Harmey, J.H.; Bucana, C.D.; Lu, W.; Byrne, A.M.; McDonnell, S.; Lynch, C.; Bouchier-Hayes, D.; Dong, Z. Lipopolysaccharide-induced metastatic growth is associated with increased angiogenesis, vascular permeability and tumor cell invasion. *Int. J. Cancer* **2002**, *101*, 415–422. [CrossRef]
58. Min, H.Y.; Lee, H.Y. Oncogene-driven metabolic alterations in cancer. *Biomol. Amp Ther.* **2018**, *26*, 45. [CrossRef]
59. Gwin, J.L.; Klein-Szanto, A.J.; Zhang, S.Y.; Agarwal, P.; Rogatko, A.; Keller, S.M. Loss of blood group antigen A in non-small cell lung cancer. *Ann. Surg. Oncol.* **1994**, *1*, 423–427. [CrossRef]