EPiC Series in Computing

Volume 77, 2021, Pages 33–40

Proceedings of ISCA 30th International Conference on Software Engineering and Data Engineering

Implementation of BERT based machine learning model to extract cancer–miRNA relationship from research literature

Arunprasad Sundharam[1] and Qin Ding[2]

[1] East Carolina University, Greenville, North Carolina, USA
sundharama19@students.ecu.edu
[2] East Carolina University, Greenville, North Carolina, USA
dingq@ecu.edu

Abstract

In the world of information technology, text mining is a widely popular methodology to extract the desired information out of the given pile of text data. Currently there are thousands of research papers/literatures published in the field of medical science related to the study of how microRNAs (miRNAs) can assist or impede the development of various types of cancer. mirCancer is a repository offered by East Carolina University to access details of cancer-miRNA association from more than 7000 research papers retrieved using rule based text mining technique. It would be a good value if we can create a machine learning model to extract the cancer-miRNA association details from the title and abstract content of these medical research papers. In this research paper, we have proposed a machine learning model which is designed and implemented using the open source NLP framework – BERT, provided by Google, to identify the cancer-miRNA relationship in the given abstract content of the research papers. We have also prepared the dataset required to train and validate the proposed model. The model developed by us performed with an overall accuracy of 90.3% in retrieving the required information from the research literatures of the test dataset and it can be useful for retrieving cancer-miRNA association information from future research literatures.

1 Introduction

MicroRNAs (miRNAs) are a class of non-coding RNAs with an average of 22 nucleotides in length and play important roles in regulating gene expression. The majority of miRNAs are transcribed from DNA sequences into primary miRNAs and processed into precursor miRNAs, and finally mature miRNAs. In most cases, miRNAs interact with the 3' untranslated region (3' UTR) of target mRNAs to induce mRNA degradation and translational repression. However, interaction of miRNAs with other regions, including the 5' UTR, coding sequence, and gene promoters, have also been reported. Under certain conditions, miRNAs can also activate translation or regulate transcription. The interaction of miRNAs with their target genes is dynamic and dependent on many factors, such as subcellular location of miRNAs, the abundancy of miRNAs and target mRNAs, and the affinity of miRNA-mRNA interactions. miRNAs can be secreted into extracellular fluids and transported to target cells via vesicles, such as exosomes,

or by binding to proteins, including Argonautes. Extracellular miRNAs function as chemical messengers to mediate cell-cell communication. MicroRNAs (miRNAs) are involved in the regulation of a variety of biological and pathological processes, including the formation and development of cancer[7]. Even though it is uncertain whether cancer is a cause or consequence of deviant miRNA expression, miRNA fingerprints are found in all types of analysed cancers, such as lung cancer, breast cancer, cervical cancer and lymphoblastic leukemia[9].

Currently there are thousands of research papers published about the association of miRNAs in various types of cancer. As the research interests on this topic keep growing, there is also need to consolidate the findings of these research papers for future works. Numerous databases have been created to document miRNA functionalities either from computational predictions or from experimental results[1]. Although computational target prediction methods are fast, experimental validation of miRNA functionalities is also needed. The significant increase in validation experiments raises the need for having a database to store these results in some uniform way[3]. However, comparing to databases providing computationally predicted miRNA functions, databases storing experimental miRNA targets are rare. There are databases which provide the details of the miRNA association on various diseases[10]. Rapid increase in the number of miRNA-related publications makes the manual collection more and more difficult[3].

miRCancer is a repository (offered by East Carolina University) where medical researchers can quickly access the microRNA–cancer associations details determined from the experimental results which are published from more than 7000 research papers. This repository was developed using rule-based text mining approach to extract miRNA and cancer association and store them in a database. All the discovered associations have been manually confirmed after automatic extraction. miRCancer initially in 2013 documented 878 relationships between 236 microRNAs and 479 human cancers through the processing of 426000 published articles from PubMed. Over the years, the repository has grown significantly and currently documents 9080 relationships between 1037 microRNAs and 131 human cancers[2].

We started this research work to pursue an alternate machine learning based approach to extract the cancer-miRNA relationship from the given abstract text of research papers, we came across the Bert framework offered by Google for NLP (Natural language Processing) tasks . We observed the capability of the Bert model can be leveraged for text mining purposes from medical research papers with proper design and training. We pursued our research to design a machine learning model using the Bert framework to extract the cancer-miRNA relationship from research papers. The following sections will explain in detail the architecture of the proposed model, training strategy of the model components and the actual results achieved using the model in extracting the cancer-miRNA relationship correctly from the medical research papers.

2 Proposed Machine Learning Model

2.1 Overview

In this research paper, we propose an alternate machine learning model which we have designed and trained to extract the cancer-miRNA relationship from the given title and abstract content of research papers using the open source Bert framework[8]. We propose the machine learning model (shown in Figure 1) to retrieve the cancer-miRNA relationship from the title and abstract content of research papers. The given abstract content along with the title of the research, (together we will refer this as input text content, hereafter) is provided as input to the proposed model.

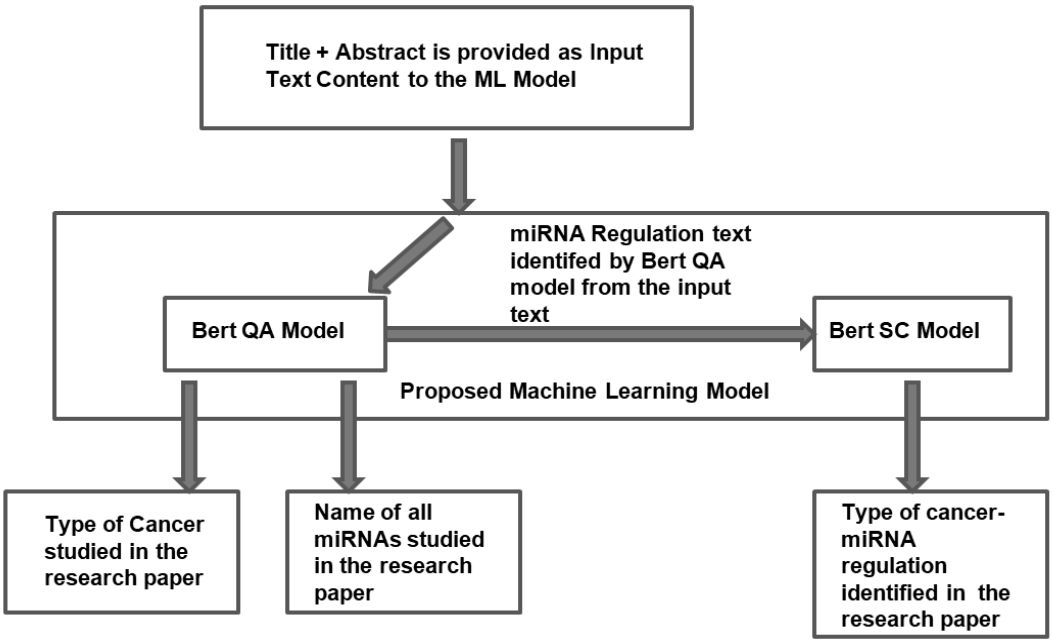

Figure 1: Proposed Machine Learning Model

The model is designed to extract only miRNA association from research papers which study on a single type of cancer. We observed only a small percentage (below 10%) of research papers study the miRNA association with more than one type of cancer. In case there is more than one type of cancer studied in the research paper, the model will retrieve only the first type of cancer mentioned in the input text content.

Also Bert does not support parsing text content of size with more than 512 tokens (as per the bert vocabulary) and hence we have excluded analyzing input text content of size with greater than 512 tokens[6]. We observed that about 222 out of the total 6422 research papers had input text content of size with more than 512 tokens which were excluded from the dataset used in this work.

To extract the cancer-miRNA association details from the input text content, the model we proposed here will retrieve the below three types of information (which we refer as information categories) from each abstract text as part of the text extraction process.

1. Type of Cancer
2. Name of the miRNA
3. miRNA Regulation

2.2 Model-Architecture

The proposed machine learning model (shown in Figure 1) consists of two component Bert Models:

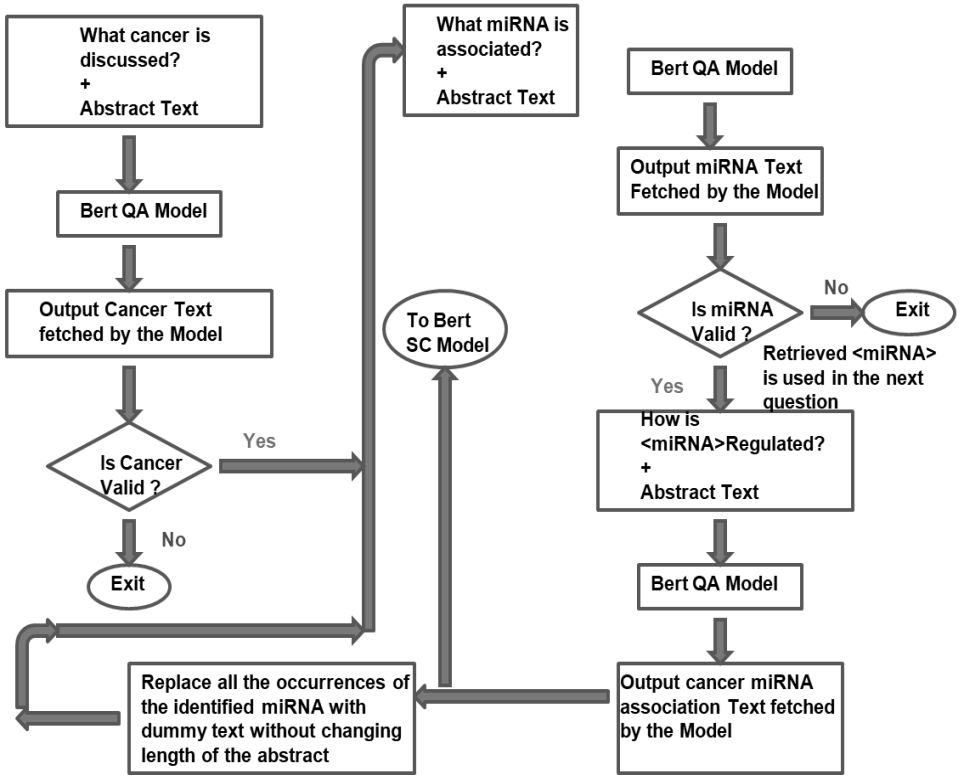

Figure 2: Bert QA Model-Training Strategy

1. The Bert QA Model which is finetuned on Question and Answering task.
2. The Bert SC Model which is finetuned on Sentence Classification task.

We used the bert base uncased model version to implement the Bert QA model and Bert SC model mentioned in the architecture.

The input text content is passed first as input to the Bert QA model which will retrieve the appropriate text from the given input for all three information types mentioned earlier. The value retrieved for the information type "miRNA Regulation" will be further passed onto the the Bert SC model where the retrieved miRNA Regulation text will be classified as UP or DOWN regulated.

2.3 Implementation of Bert QA Model

The Bert QA (Question Answering) model used in the proposed model is implemented [4] using the training strategy described in Figure 2.

1. The input text content is first concatenated with the question text "What cancer is discussed?". The Bert QA model analyzes the concatenated text and retrieves the type of cancer studied in the input text content. If there is no valid type of cancer available in the text, the model skips the remaining steps and exit or move to the next input text content.

2. If a valid type of cancer is available in the given input text content, the input text content is then concatenated with the question text "What miRNA are associated?".The Bert QA model analyzes the concatenated text and retrieves the first miRNA name mentioned in the input text content. If there is no valid miRNA name available in the given text, the model skips the remaining steps and exit or move to the next input text content

3. We will use the identified miRNA in place of text in bold **miRNA** in the following question text -"How is **miRNA** regulated?". The abstract text is then concatenated with this question. The Bert QA model analyzes the concatenated text and retrieves the text related to how the cancer is associated by the identified miRNA. The retrieved text is then passed onto the Bert SC model to classify it as either 'UP' or 'DOWN' regulated.

4. In this step, we will replace all the occurrences of the retrieved miRNA name in the given input text content with the dummy text without changing the length of the input text content. We achieve this by creating a dummy text exactly in same length of the identified miRNA text length by concatenating the text "miRNA" followed by the required number of alphabet 's' to match the same length of the identified miRNA name.

   For example, if the identified miRNA text is 'mir-7-5p' which is of 8 characters in length, we replace all the occurrences of 'mir-7-5p' in the abstract text with 'miRNA' followed by 3 's' (.i.e. miRNAsss).

5. Now we repeat the Steps 2-4 to find the next unique miRNA name mentioned in the given input text content until there is no valid miRNA name is available in the updated input text content

2.4 Implementation of Bert SC Model

We implemented a Bert SC (sentence classification) model [5] to classify the miRNA regulation text output retrieved by the Bert QA model from the abstract text as either 'UP' or 'DOWN' regulated. The Bert SC model is implemented using the training strategy described in Figure 3. The miRNA regulation text is first concatenated with the question "How is ¡miRNA¿ regulated?". It should be noted that the actual miRNA identifed by the Bert QA model corresponding to the miRNA regulation text is populated in place of the term "¡miRNA¿" in the above question. The Bert SC model classifies the concatenated text as either '1' or '0'. This value is converted to 'UP' or 'DOWN' as per the below rules:

- If the miRNA regulation text is classified as '1' by the Bert SC model, we determine the cancer is 'UP' regulated by the miRNA.

- If the miRNA regulation text is classified as '0' by the Bert SC model, we determine the cancer is DOWN' regulated by the miRNA.

2.5 Categorization of training and test dataset

We gathered the input text content (title + abstract) and cancer-miRNA association details of 6422 research papers from the miRCancer repository (http://mircancer.ecu.edu). 13 research papers did not have abstracts available as these articles were retracted. We randomly split the gathered details of the available 6409 research papers into different two sets.

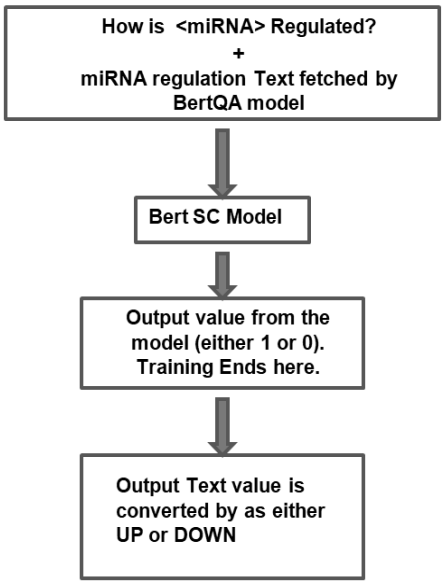

Figure 3: Bert SC Model-Training Strategy

1. **Training Dataset**: We created a corpus consisting of input text content (title + abstract) and cancer-miRNA association details from 5197 research papers. 188 research papers had input text content of size with more than 512 tokens (words in Bert vocabulary) limit and hence were excluded from further processing. Also 330 research papers did not contain the expected cancer-miRNA relationship in the provided text during manual verification and were excluded. We finally had a corpus consisting of input text content and and cancer-miRNA association details from 4679 research papers which was used for the creation of dataset required to train the model.

2. **Test Dataset**: We created another corpus consisting of input text content (title + abstract) and cancer-miRNA association details from 1212 research papers. 34 research papers had input text content of size with more than 512 tokens (words in Bert vocabulary) limit and hence were excluded from further processing. Also 33 research papers did not contain the expected cancer-miRNA relationship in the provided text during manual verification and were excluded. We finally had a corpus consisting of input text content and cancer-miRNA association details from 1145 research papers which was used for the creation of dataset required to validate/test the model.

3 Results

We validated the implemented machine learning model using the test dataset (consisting of input text content from 1145 research paper abstracts). Table 1 (Validation results of the model against test dataset) shows the details of the results observed from the validation of the model against test dataset. The percentage of correct predictions for the three information

Metrics from validation results	**Value**
Total number of cancer available in the test dataset	1145
Number of correct cancer predictions	1131
Total number of miRNA available in the test dataset	1213
Number of correct miRNA predictions	1195
Total number of miRNA regulation available in the test dataset	1213
Number of correct miRNA regulation predictions	1107
Total number of predictions made by the model in the test dataset	3571
Total number of correct Predictions	3433
Percentage of correct cancer predictions	98.78%
Percentage of correct miRNA predictions	98.52%
Percentage of correct miRNA regulation predictions	91.26%
Percentage of total correct predictions	96.14%
Number of research papers in the test dataset with cancer-mirna association available	1145
Number of research papers in the test dataset with all 3 information categories predicted correctly	1035
Percentage of research papers with correct predictions for cancer-mirna association	**90.39%**

Table 1: Validation Results of the model against test dataset

categories- type of cancer, name of the miRNA and miRNA Regulation are observed to be 98.78%, 98.52% and 91.26% respectively. The overall percentage of correct predictions of the model is observed to be 96.14%. But we can consider the cancer-miRNA association predicted by the model for a research paper is correct, only when the predictions of the model is correct for all the three information categories of that particular paper. We validated the prediction results of the model to determine the number of research papers for which the model predicted correct values for all the three information categories. We observed the model predicted the cancer-miRNA association details correctly for 90.39% of the research papers in the test dataset.

4 Conclusion

In this paper, we have designed and implemented a machine learning model using Bert framework to extract the information of cancer-miRNA association from the research papers. During the course of this work, we also prepared the dataset required to train the model in the task of identifying cancer-miRNA relationship from the given text. The machine learning model implemented in this work is observed to perform with prediction accuracy of 90.39% (against test dataset). It can serve as a quick and effective means of extracting cancer-miRNA association details from the research papers. As the model is implemented and trained directly from the abstract text, it will learn to accomodate any future texts as it will continuously train the model using future literatures and it does not require any manual analysis of the text to add new rules. In this work, we have excluded research papers with input text content (title + abstract text) of size greater than 512 tokens as it can not be parsed by the Bert model. Also the model in its currently implemented form is designed to handle research literatures in which

only single type of cancer is studied. If there is more than one type of cancer studied in the research paper, the model will retrieve only the first type of cancer mentioned in the input text content.We can extend this model to handle multiple types of cancer mentioned in the given abstract as part of future scope of work on this topic. The work done in this paper to generate the training and test dataset could be of good use for future enhancements in this area.

References

[1] Boya Xie, Qin Ding, Di Wu. Text mining on big and complex biomedical literature. http://www.ctan.org/tex-archive/help/Catalogue/entries/inputenc.html, last viewed April 2010, 1986–2009.

[2] Boya Xie, Qin Ding, Hongjin Han, and Di Wu. mirCancer Repository archive network. http://mircancer.ecu.edu/, 2012.

[3] Boya Xie, Qin Ding, Hongjin Han, Di Wu. mircancer: a microrna–cancer association database constructed by text mining on literature. https://doi.org/10.1093/bioinformatics/btt014, 2013.

[4] Hugging Face. Bert for question answering. https://huggingface.co/transformers/model_doc/bert.html#bertforquestionanswering, 2019.

[5] Hugging Face. Bert for sequence classification. https://huggingface.co/transformers/model_doc/bert.html#bertforsequenceclassification, 2019.

[6] Google. Bidirectional Encoder Representations. https://github.com/google-research/bert, 2018.

[7] Hwang,H.W. and Mendell,J.T. Micrornas in cell proliferation, cell death, and tumorigenesis. https://pubmed.ncbi.nlm.nih.gov/16495913/, 2006.

[8] Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805, 2018.

[9] Jacob O'Brien, Heyam Hayder, Yara Zayed, Chun Peng. Overview of microrna biogenesis, mechanisms of actions, and circulation. https://www.frontiersin.org/articles/10.3389/fendo.2018.00402/full, 2018.

[10] Q1 Jiang. mir2disease: a manually curated database for microrna deregulation in human disease. https://www.researchgate.net/publication/23388679_miR2Disease_a_manually_curated_database_for_microRNA_deregulation_in_human_disease, 2008.