Transfer Learning in Classifying Prescriptions and Keywords-based Medical Notes

Mir Moynuddin Ahmed Shibly
Dept. of Computer Science & Engineering
East West University
Dhaka, Bangladesh
shiblygnr@gmail.com

Tahmina Akter Tisha
Dept. of Computer Science & Engineering
East West University
Dhaka, Bangladesh
tahminatish001@gmail.com

Md. Kamrul Islam
Dept. of Computer Science & Engineering
East West University
Dhaka, Bangladesh
kamrulewucse@gmail.com

Abstract – Medical text classification is one of the primary steps of health care automation. In this study, two types of medical texts were classified into some medical specialties. The first one is the keywords-based medical notes and the second one being prescriptions. The objective of this study was to analyze the prospects of transfer learning in medical text classification. To do that, a transfer learning system was created for classification tasks by fine-tuning Bidirectional Encoder Representations from Transformers aka BERT language model and its performance was compared with three deep learning models – multi-layer perceptron, long short-term memory and convolutional neural network. The fine-tuned BERT model showed the best performance among all the other models in both classification tasks. It had 0.84 and 0.96 weighted f1-score in classifying medical notes and prescriptions respectively. This study proved that transfer learning can be used in natural language processing, and significant improvement in performance can be achieved through it.

Keywords – Transfer learning, BERT, natural language processing, medical text classification, medical notes, prescriptions

I. INTRODUCTION

Text classification is one of the major areas of natural language processing. It has many applications in fields like sentiment analysis, intent recognition, spam filtering, etc. An automated system to classify texts can help us in several ways. Every application will eventually need machine learning-powered artificially intelligent natural language processor incorporated in it to provide more user-friendly experiences. Machine learning-based AI systems can act as decision support systems too. An AI support system can add more efficiency in decision making [1] in various fields. A decision support system was developed to maintain the electrical grid [2] another one was created to maintain coordination of product design [3] and so on. There are many examples of such decision-making support systems. The necessity of such support systems in the medical sector is immense. This kind of support system can assist a doctor to classify the patients into some appropriate groups so that the doctor can treat them better. An appropriate system can guide a patient to better diagnosis as well. Moreover, The implementation of these types of support systems will ensure more automation in the health care sector to provide better services.

A. Background and Problem Statement

There are many studies related to the detection of medical specialties, domains, and diseases using machine learning approaches. Some of them are text-based classification, and some are number-based. In a study [4] conducted in Germany, a number-based classification system was developed to predict the next events for patients whose kidney was transplanted based on the static attributes like gender, age, and dynamic attributes like medical reports, medications. The model was created using a recurrent neural network with the gated recurrent unit and predicted one of the three outcomes – kidney rejection, kidney loss and death of the patient. The number-based classification task is good but text-based classifications are more appropriate to detect a specific medical sub-domain or specialty or next event in the early stage of diagnosis.

Trang Pham at el. [5] had used natural language processing to predict disease progression, intervention recommendation, and future risk in two medical specialties – diabetes and mental health. Long short-term memory (LSTM) with RNN had been modified as care-LSTM in completing those works. As LSTM architecture can hold the contextual information of a sequence by passing only the relevant information to the forward section of the neural network, it is widely used in classifying texts of the medical domain. Other neural network architectures like multi-layer perceptron and convolutional neural network can also be used to process clinical texts.

Another promising technique to classify something is transfer learning. Transfer learning is widely used in various types of classification tasks. Biomedical domains are not different from that. It had been used for automated plant identification systems from images that helped stakeholders to deal with the enormous number of plants quickly by mitigating the required time and expense of these operations. The use of transfer learning improved the models for the classification of plants with low performance. It has helped to more accurately identify plant species, which has many benefits not only in agriculture but also in the field of biodiversity, health, and forest studies [6]. It had also been used in detecting breast cancer [7]. Apparently, maximum works done in the health sector using transfer learning are image-based. There are very few studies that had been conducted to classify medical texts using transfer learning. This study aims at exploring that gap of knowledge. In this study, two types of medical documents – prescriptions and medical notes are going to be classified using transfer learning.

B. Aim and Objectives

This study explores the prospects of transfer learning models using natural language processing to classify medical prescriptions and medical notes into some medical specialties.

1. To design neural network-based models to classify keywords-based medical notes and medical prescriptions.
2. To design transfer learning models to classify keywords-based medical notes and medical prescriptions.
3. To analyze both types of models and compare their results.

II. Related Works

A study [8] showed that even with a larger number of parameters, RNN models like LSTM and GRU had given a good performance on smaller dataset sizes to detect medical events from the noisy natural text of electronic health record (EHR) notes. The authors claimed themselves to be the first group in reporting the use of RNN frameworks for information extraction from EHR notes. The authors evaluated both methods on 780 EHR notes of hematological malignancy cancer patients empirically.

The authors in the paper [9] had constructed a pipeline using a clinical natural language processing system. They had created a medical subdomain classifier from two datasets - Integrating Data for Analysis, Anonymization, and Sharing (iDASH) data repository and Massachusetts General Hospital (MGH) dataset using 15 different combinations of data representations methods and supervised machine learning algorithms. First, the baseline model had been created using seven shallow machine learning algorithms. Then the best shallow machine learning model was compared with two models created with convolutional neural network and convolutional recurrent neural network. The CRNN model had outperformed all the other models.

Another research had been carried out to model a system to diagnose appendicitis based on clinical notes [10]. The researchers in this study had proposed a neural network model combining CNN, recurrent neural network and residual neural network to detect whether a patient has appendicitis or not based on the notes created by emergency doctors. LSTM technique had also been used to control the flow of information. The created model can predict up to 90% accurately about a patient having appendicitis or not while ED doctors can predict up to 89.5%.

Transfer learning had been used in biomedical natural language processing [11]. To simplify the research in the development of pre-training language representation of the biomedical domain, a benchmark had been designed by the authors. They pre-trained the BERT model with PubMed abstract and MIMIC III datasets. Transfer learning was also applied in recognizing named entity on Chinese medical records [12]. Transfer learning was combined with bi-directional LSTM to improve performance on NER in electronic medical records.

From the analysis of the works related to this study, it shows that deep neural network-based approaches can be used in both of the classification tasks of this work. And though it was not used widely to classify clinical texts, transfer learning has good prospects in this matter. This scientific base leads to the following research question:

What are the prospects of using transfer learning in classifying medical notes and prescriptions into the specific domain?

III. Dataset

The medical prescriptions and medical notes containing keywords had been classified based on two publicly available datasets – MTSamples clinical notes dataset[1] and prescription dataset[2]. There were transcribed data about 40 medical specialties in the MTSamples dataset. And the prescription dataset contained prescribed medicine of 282 specialties. Each instance of both datasets was labeled accordingly with a single medical specialty. Five specialties from the clinical notes dataset and seven specialties from the prescription dataset were chosen based on their unique features for the classification tasks in this work.

IV. Methodology

Three approaches were followed to classify the keywords based on medical notes and prescriptions. First of them was the classical machine learning approach, the second one was a deep neural network approach. And the final approach to categorize the medical notes and the prescription was to use transfer learning by fine-tuning Google's Bidirectional Encoder Representations from Transformers (BERT) model.

A. Classical Machine Learning Approach

Three of the conventional machine algorithms were applied in this study – Naïve Bayes, Decision Tree, and Random Forest classifiers.

1. Text Pre-processing

Before applying the classifiers, the text data of medical notes and prescriptions were preprocessed. Initially, the data were cleaned by removing HTML tags, English stop words, any undesired character, and numbers. Then, a dictionary with all the vocabularies of training instances was created following an indexing algorithm, and each instance of the training examples was vectorized with the help of the created dictionary. The length of each vector was the same as the length of the dictionary. Basically, the whole training set was converted to a sparse matrix where each row represented each instance of the training set. When the preprocessing was done, the training set was fed into the created machine learning models.

2. Algorithms

Naïve Bayes classifier is one of the most popular classifiers in the area of text categorization [13]. It is a probabilistic classifier that takes a naive approach to the features being independent of each other to the basic Bayesian model. One of the baseline-model was created using this naïve assumption. Another classical machine learning classifier is *decision tree*. It continuously divides the work area by plotting lines into sub-areas until a specific class emerges. The Iterative Dichotomiser 3 (ID3) decision tree learning algorithm [14] was used considering all the features of the dataset as discrete values. The final baseline-model for this study was created by the *random forest* algorithm. It is an ensemble learning method which is widely used in regression and classification problems. A random forest is nothing but a combination of many decision trees. It resolves the problem of overfitting in decision tree classifier [15].

B. Deep Learning Approach

In this phase of the study, medical notes and prescriptions were classified using deep learning techniques. Three of the neural network approaches were applied – multi-layer perceptron, a recurrent neural network with bi-directional long short-term memory and convolutional neural network.

[1] www.mtsamples.com

[2] www.kaggle.com/roamresearch/prescriptionbasedprediction

1. Text Pre-processing

The text preprocessing step of feeding the neural network was different than those of the classical machine learning algorithms. For the classical approach, the bag of words (BOW) technique was used to make the text input ready by converting them into a sparse matrix. In contrast, at this stage of deep learning approach, after cleaning the dataset, a dictionary was created which contained each unique word of the training examples with an integer mapped with it. Then each training example was converted to a fixed-length sequence of mapped numbers from that dictionary. If an instance had fewer word numbers than the fixed length, the rest of the sequence was padded with zero, and when the number is longer, the overflowed words were truncated as the length of all training examples had to be the same. For the testing instances, when any word was found which was not present in the dictionary, it was replaced with an arbitrary token. That token was the same for all the testing examples.

2. Multi-layer Perceptron

Multi-layer perceptron is a feed forward artificial neural network consists of at least three layers – one input layer, one output layer, and a minimum one hidden layer. The number of hidden layers varies depending on the complexity of the problem that needs to be solved. A multi-layer perceptron generally works in three phases. First, the inputs keep propagating forward from the input layer to the output layer by passing through each hidden layer. In every hidden layer, the weights get updated by multiplying the previous weights with inputs and adding bias to it. The prediction is made after completing each forward pass of the input to the output layer, and the loss is determined based on the learning rate. Finally, the loss is propagated backward, and weights are eventually updated at each hidden layer. The layers use some activation functions that map the inputs to the output.

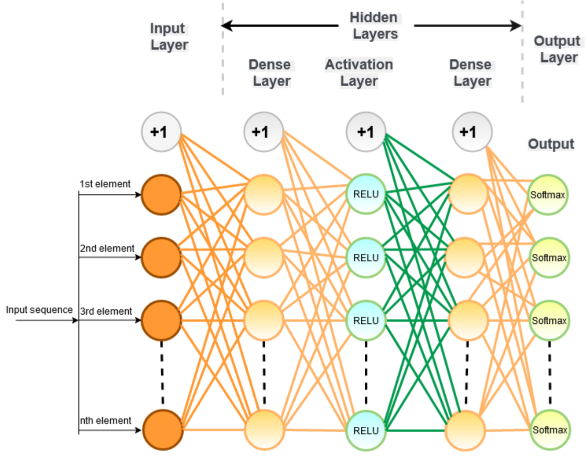

Fig 1: MLP architecture for classification

Here, a simple 5-layer feed forward artificial neural network had been used for classification challenge. The layers are- an input layer having a fixed-length number of neurons, a dense layer, an activation layer with rectifier linear unit (RELU) (i) activation function, another dense layer, and the output layer with softmax (ii) activation function. The output layer had 5 neurons for the medical notes dataset and 7 neurons for the prescription dataset based on the number of target classes.

$$f(x) = x^{+} = \max(0, x) \ldots\ldots..\text{eq(i)(relu activation)}$$

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \ldots\ldots..\text{eq(ii)(softmax activation)}$$

3. Long Short-Term Memory

A recurrent neural network (RNN) is a class of artificial neural networks where node-to-node connections form a directional graph along a time continuum. Long short-term memory adds feedback connections to the regular RNNs so that it can predict something based on only relevant information that carried down the network. Regular RNNs face vanishing gradient problem in the early layers of the network during back-propagation [16]. LSTM architecture prevents RNN from that. LSTM architecture consists of a cell and three gates – input, output and forgets gate. The cell keeps track of the dependencies among the input sequence that can lead this architecture to perform better than other systems. In this system, a forget gate determines which information are to keep, and which are to forget. After passing the forget gate, there comes an input gate which updates the cell state. And, as well an output gate to pass the information to the next hidden state. Those gates have sigmoid activation function which limits the values between 0 and 1.

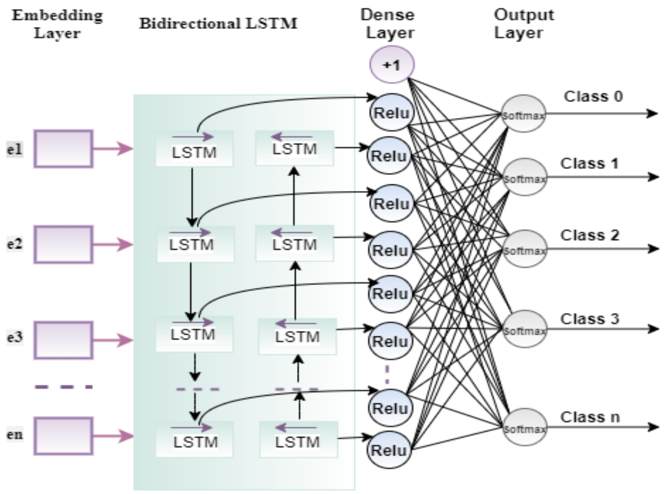

Fig 2: LSTM architecture

In this work, a bi-directional LSTM layer had been used after an embedding layer. The embedding layer is the first layer of the network which takes the word embeddings as input. After the bi-directional layer, there were two dense layers having RELU and softmax activation functions respectively.

4. Convolutional Neural Network

Though the convolutional neural network (CNN) is generally used for image classification, it also gives good performance when comes to the task of classifying texts. An eight-layer CNN had also been used in this study.

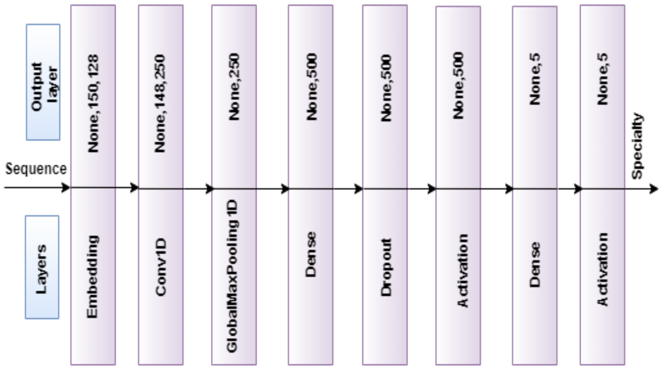

Fig 3: CNN architecture for notes classification

C. Transfer Learning

Transfer learning is the method of transferring knowledge of solving a problem to solve another related problem [17]. It has shown significant improvement in performance for image classification and computer vision. Using a pre-trained model, images can be classified in desired categories having more accuracy with a relatively smaller dataset. The prospects of transfer learning in natural language processing has been recognized when Google developed their language model Bidirectional Encoder Representations from Transformers – BERT [18]. It was a breakthrough in the field of natural language processing. There are a few other pre-trained models – ELMo [19], ULMFiT [20]. In this study, the BERT model was fine-tuned to do the classification tasks of the medical domain. Using it, there is no need for designing new layers to train these domain-specific data. Instead, the model can be created by utilizing the pre-trained model with a slight modification to reach the stat-of-the-art goal in text classification.

1. Fine-tuning with BERT

There are many versions of BERT models[3]. The BERT base version had been fine-tuned to classify the prescriptions and medical notes in this work. It is a 12-layer network with 768 hidden dimensions, and it was pre-trained on 110 million parameters.

To use this pre-trained model, first comes the stage of text pre-processing. There are many words in two datasets of this work like medicine name, test name, medical abbreviations, etc. that are not present in the vocabulary of the pre-trained BERT. Despite the actual words not being present in the vocabulary, BERT has some partially filled words in it. It uses those words to map the unknown words. For doing this, a single word can be decomposed in a few different words. The reason behind the success of this technique is that a certain word gets decomposed in the same way throughout the entire working process. Therefore, all the instances having that particular word have the same impact. Here is an example of how BERT handles unknown words:

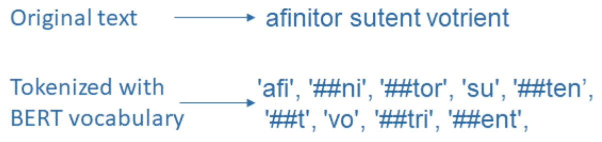
Original text
afinitor sutent votrient
Tokenized with BERT vocabulary
'afi', '##ni', '##tor', 'su', '##ten', '##t', 'vo', '##tri', '##ent',

Fig 4: Handling unknown words by BERT

Since a single word can be decomposed in many, another problem arises. The input layer of the pre-trained BERT model has a maximum sequence length of 512. Any sequence larger than 512 will cause an error as the BERT model simply cannot process it. Due to the splitting, a sequence may become longer than 512. At that point, the sequences after 512 have to be truncated and some information loss occurs. The fine-tuned model had an input layer before the sequences entered the BERT model. And after the BERT model, two layers had been added – naming, dropout, dense with RELU activation function, dropout and dense with softmax activation layers.

Explaining how the BERT model processes data is beyond the scope of this study. The original BERT study [18] is recommended for understanding the internal mechanisms of this pre-trained model. In this way, adding BERT pre-trained model in between task-specific input layer and output layers allows the input sequences to go through all those BERT layers. It adds an extra dimension to accomplish the specific tasks that the model has been created to do – in this case, classifying medical prescriptions and notes.

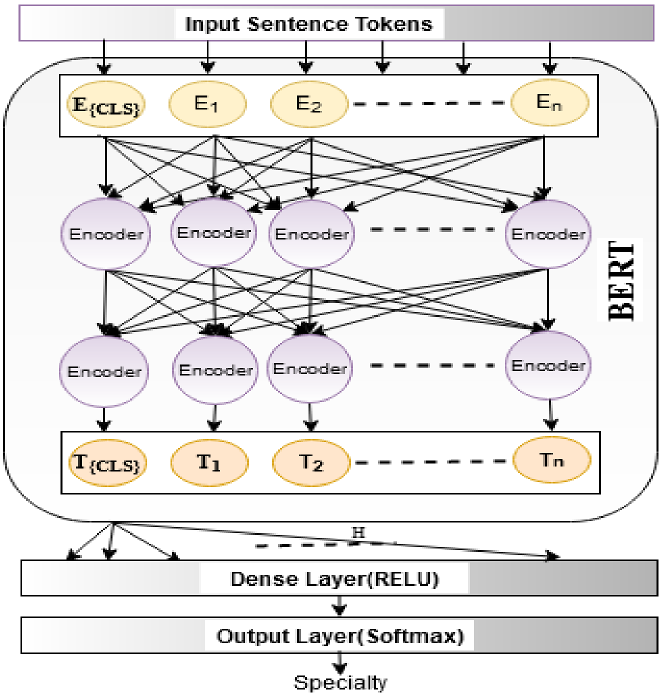

Fig 5: Fine-tuned BERT model

V. Experimental Setup and Results

The aim of this study is to classify medical notes and prescriptions into some medical domains. Each of the transcribed data from the MTSamples medical notes dataset has various information about a patient like the history of past illness, current symptoms, procedure description of treatment, keywords, etc. And, each of the notes is labeled with one of those 40 medical specialties. The scope of this study was limited only to the keywords of the notes. Hence, the first task was to extract those keywords and store them into a CSV file with an appropriate label. 5 out of those 40 medical specialties were selected for this study – Gastroenterology, Neurology, Orthopedic, Radiology, and Urology. The reason behind choosing these five categories was that these specialties had a balanced number of examples, while some other specialties had very few, and others had too many. The keywords of these five medical specialties were extracted manually from the MTSamples website, and each instance was labeled appropriately. Some of the instances had the medical specialty itself in the keyword list. Those were removed to eliminate bias from the dataset.

On the other hand, the prescription dataset had the name of medicines prescribed by the specialists. This dataset had 282 medical specialties, and among those, 7 specialties were chosen for the classification task. The selected specialties are – Cardiovascular Disease, Gastroenterology, General Practice, Hematology and Oncology, Nephrology, Neurology, and Psychiatry.

[3] https://github.com/google-research/bert

For the prescription dataset, all instances of the selected specialties were not chosen. Rather, 300 instances of each class were selected randomly to make sure the balance of data. After selecting data for classifying, those were pre-processed and made ready for input. The preprocessing mechanism was described in the methodology section.

The data were split into two sets initially – training and testing. The training set had 80% and the testing set had 20% data. The vocabulary was created only based on the training set. The testing set was completely isolated from the training phases. For deep and transfer learning models, another 10% data were used for validation purpose. The experiment of this study was divided into three phases – creating baseline models using classical machine learning algorithms, creating deep learning models and transferring knowledge of the BERT language model into the medical domain specific classification. In this section, the results of the three phases are presented and the results of each phase were compared. In this work, four evaluation metrics had been used – precision, recall, f1-score and accuracy to measure the performances of the created models.

Baseline models and deep learning models were trained on a computer having Ryzen 5, 1600 CPU and Nvidia GeForce 1050TI GPU with Linux Mint 19.3 operating system. And, the fine-tuned BERT model was trained on colab.research.google.com with its tensor processing unit (TPU) support.

1. Baseline Models

Initially, a baseline had been created for both classification tasks using classical machine learning algorithms – decision tree, random forest, and naïve Bayes classifiers. The scikit-learn[4] library of python was used to create baseline models for classifying the keywords into 5 medical specialties and to classify the medical prescriptions into 7 specialties. Among the classical machine learning models for classifying keywords, the performance of the decision tree model was best. It had a weighted f1-score of 0.74 and 0.74 accuracy. On the other hand, in classifying prescriptions, both Naïve Bayes and Random Forest classifiers had a similar good performance. Both classifiers had a weighted f1-score of 0.93 and 0.93 accuracy.

2. Deep Learning Methods

In the second step of the experiment, three deep learning models – multi-layer perceptron, LSTM, and the convolutional neural network had been created using Keras on top of TensorFlow[5] library of python. All the deep learning models demonstrated better performances in notes classification than the baseline models having a weighted average f1-score of 0.76, 0.80 and 0.79 respectively. While they showed 0.95, 0.95 and 0.95 weighted average f1-scores in classifying prescriptions into seven categories. In both cases, LSTM models had the best results amongst the deep learning models. Specialty-wise precision, recall, and f1-score of the best deep learning model for both datasets are visualized in fig 6 and 7. While training these deep learning models, various optimizers like Adam, SGD, Adagrad, Adadelta, RMSprop, etc. were used, and Adam optimizer showed the best result in both classification tasks for all three models.

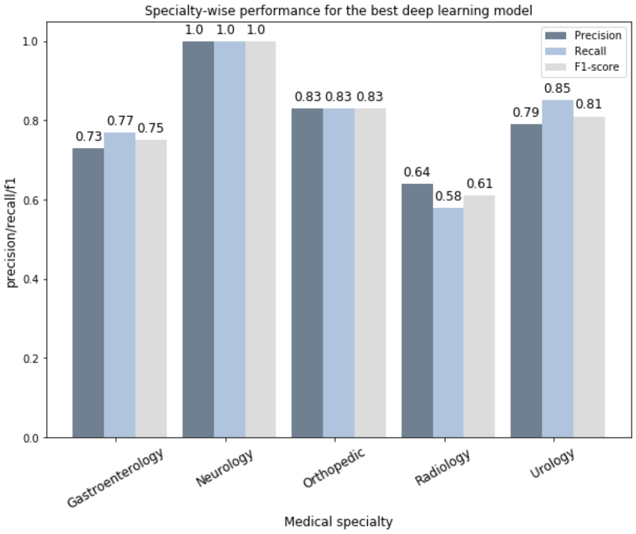

Fig 6: Specialty-wise performance of the best deep learning model (notes)

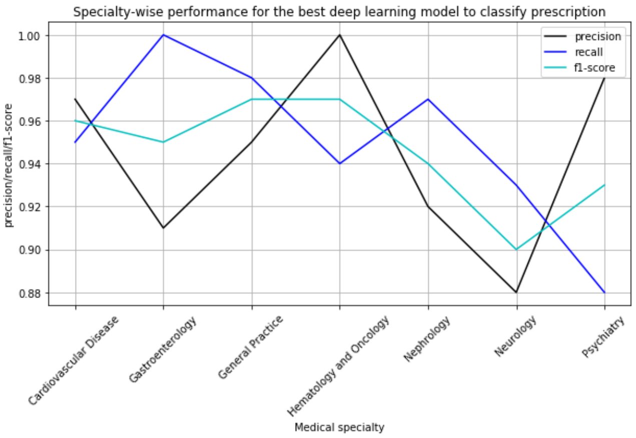

Fig 7: Specialty-wise performance of the best deep learning model (prescription)

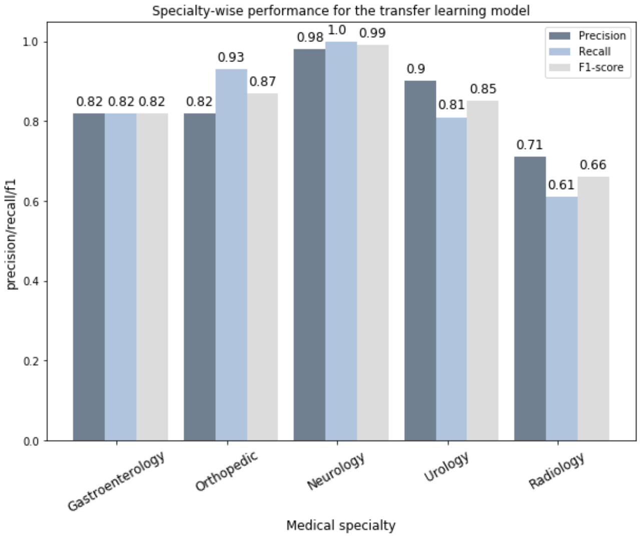

Fig 8: Specialty-wise performance of the transfer learning model (notes)

[4] https://scikit-learn.org

[5] https://www.tensorflow.org

Dataset		Baseline			Deep Learning			Transfer Learning with BERT
		Decision Tree	Naïve Bayes	Random Forest	Multi-layer Perceptron	Long short-term memory	Convolutional neural network	
Notes	Precision	**0.74**	0.71	0.72	0.76	**0.80**	0.79	**0.84**
	Recall	**0.74**	0.70	0.71	0.77	**0.80**	0.79	**0.84**
	F1-score	**0.74**	0.70	0.71	0.76	**0.80**	0.79	**0.84**
	Accuracy	**0.74**	0.70	0.71	0.77	**0.80**	0.79	**0.84**
Prescription	Precision	0.88	**0.94**	**0.94**	0.95	**0.95**	**0.95**	**0.97**
	Recall	0.87	**0.93**	**0.93**	0.95	**0.95**	**0.95**	**0.96**
	F1-score	0.87	**0.93**	**0.93**	0.95	**0.95**	**0.95**	**0.96**
	Accuracy	0.87	**0.93**	**0.93**	0.95	**0.95**	**0.95**	**0.96**

Table 1: Performances of all created models

3. Transfer Learning with BERT

Lastly, in the final step, state-of-the-art Google-developed language model – BERT was fine-tuned for the classification tasks. The transfer learning models had also been trained using TensorFlow. The BERT model outperformed the baseline models and the other deep learning models as well. It had a weighted average f1-score of 0.84 in classifying keywords based medical notes. On the other hand, it had 0.96 f1-score in classifying the other task. Category wise performance of the fine-tuned BERT model is shown in fig 8 and fig 9 – the first one having the visualization of how the model performed for each class of the first dataset in terms of precision, recall, and f1-score, and the second one having the performance visualization of the model in the second dataset.

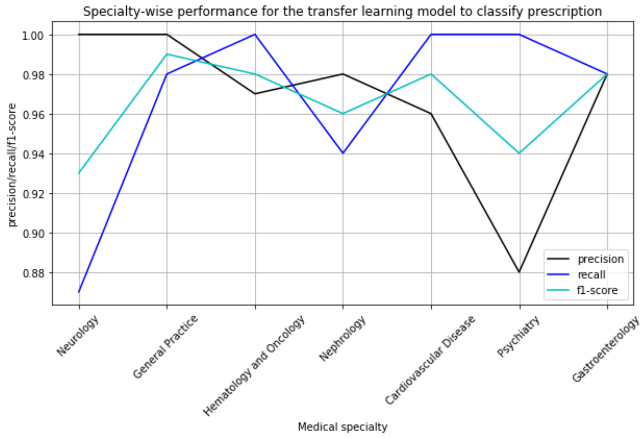

Fig 9: Specialty-wise performance of the transfer learning model (prescription)

VI. Discussion

This study demonstrates how various methods can be used for the classification tasks of the medical domain using natural language processing. These types of works are very important to provide a meaningful support system for the doctors in order to smooth their workflow. A machine learning-based system can help a doctor in so many ways such as treatment, disease prediction, etc. Moreover, patients find it difficult to consult the exact medical specialists based on their medical problems in many cases. Focusing on this problem, the study aims to efficiently classify two types of documents corresponding to a patient in some medical specialty, so that he/she does not have to face any difficulty to find the right specialist according to his/her medications and symptoms. According to the objectives of this study, three different types of machine learning models were built to classify keyword-based medical notes and medical prescriptions into some medical specialties. Before that, the key attributes to classify them were identified. Then, the results were compared with each other. The overall result was good for prescription classification than keyword classification. However, the performance on the keyword classification was not that bad. There were some similar studies like this. In a study [9], the researchers had classified the medical notes into six medical sub-domains while Jingshu Liu et al. [21] had predicted three chronic diseases based on clinical notes. Transfer learning had been used in the medical sector to recognize named entities [12] [11]. The performance of each model was measured based on precision, recall, f1-score, and accuracy. Table 1 shows the complete result of this study from the performance of baseline models to the performance of transfer learning models.

Among the three deep learning models, the BiLSTM model performed better than CNN and MLP models in both cases. Because only LSTM holds the contextual information about the texts. The convolutional neural network also showed good results in classifying texts, despite being widely used in image classification. By looking at the table, it becomes apparent that the fine-tuned BERT model showed the best result among all the models. The result of this study showed that transferring knowledge from one model to another domain specific task in NLP can be really useful. Transfer learning in image processing is going on for a while. But, in NLP, the idea of transfer learning is new, and it is emerging at great speed after the development of Google's BERT language model. The prospects of this type of learning are very promising, even though the domain-specific tasks differ from the original pre-trained model. In this study, fine-tuning on the pre-trained model for classifying notes and prescription yielded the best result over the baseline and regular deep learning model despite the original model was trained on very different data than the medical sector.

Only keywords from medical notes were classified in this study. The full medical notes with the past history of illness and with the procedure of treatment were not classified due to the excessive length of the texts. As mentioned earlier that the pre-trained BERT model only can accept the input sequence having a length of less than or equal to 512. The majority of those clinical notes had more than 1000 sequences. If the classification of those lossy input sequences would have done, that would not have been a proper classification. All of the medical specialties were not considered while classifying due to an imbalance in the dataset. Moreover, only an Adam optimizer was used while training the transfer learning model. The performance of other optimizers in transfer learning could not be measured due to the limitation of computational resources.

VII. Future Work

This study explored the usability of transfer learning in text classification. It also proved the acceptance of deep learning models in the medical domain-specific classification tasks. Only five to seven medical specialties were considered in this study. Adding more specialties would add robustness to this type of machine learning-based support system. More specialty-wise data could be collected to enhance performance. Another area to explore is classifying the complete clinical notes including keywords, symptoms, history of the patients, etc. Recommending medications based on the diseases using transfer and deep learning models is another promising field worth looking at. The majority of clinical notes which are available for research purposes are mainly unstructured. Preparing them in a structured way can definitely help the researchers to get the most out of these notes. Moreover, applying transfer learning in more sustainable datasets like MIMIC III, Massachusetts general hospital clinical notes dataset, n2c2 clinical NLP challenges dataset can produce good outcomes. Furthermore, using medical domain-specific language models and creating them would change the landscapes of automation in the medical sector.

VIII. Conclusion

In this study, the prospects of using transfer learning in classifying keyword-based medical notes and medical prescriptions had been explored. And, transfer learning had shown great improvement in performance over the other baseline and deep learning models which was created by fine-tuning BERT model. The transfer learning model had better performance than the recurrent bi-LSTM, MLP and CNN models. Despite the original model not being trained on the medical language, the fine-tuned model was able to demonstrate better performance because of all the pre-learned weights it had. Moreover, this study essentially proved the impact of transfer learning on text classification by achieving improved performance.

References

[1] G. Phillips-Wren, "AI tools in decision making support systems: A review," *Int. J. Artif. Intell. Tools*, vol. 21, no. 2, 2012, doi: 10.1142/S0218213012400052.

[2] C. Rudin *et al.*, "Machine learning for the New York City power grid," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 328–345, 2012, doi: 10.1109/TPAMI.2011.108.

[3] N. Taghezout and P. Zaraté, "An agent-based simulation approach in an IDSS for evaluating performance in flow-shop manufacturing system," *Intell. Decis. Technol.*, vol. 5, no. 3, pp. 273–293, 2011, doi: 10.3233/IDT-2011-0111.

[4] C. Esteban, O. Staeck, S. Baier, Y. Yang, and V. Tresp, "Predicting Clinical Events by Combining Static and Dynamic Information Using Recurrent Neural Networks," *Proc. - 2016 IEEE Int. Conf. Healthc. Informatics, ICHI 2016*, pp. 93–101, 2016, doi: 10.1109/ICHI.2016.16.

[5] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *J. Biomed. Inform.*, vol. 69, pp. 218–229, 2017, doi: 10.1016/j.jbi.2017.04.001.

[6] A. Kaya, A. S. Keceli, C. Catal, H. Y. Yalic, H. Temucin, and B. Tekinerdogan, "Analysis of transfer learning for deep neural network based plant classification models," *Comput. Electron. Agric.*, vol. 158, no. January, pp. 20–29, 2019, doi: 10.1016/j.compag.2019.01.041.

[7] S. H. Ripon, F. H. Proma, and F. Khan, "4.1 introduction," pp. 59–85, 2019.

[8] A. N. Jagannatha and H. Yu, "Bidirectional RNN for medical event detection in electronic health records," *2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. NAACL HLT 2016 - Proc. Conf.*, pp. 473–482, 2016, doi: 10.18653/v1/n16-1056.

[9] W. H. Weng, K. B. Wagholikar, A. T. McCray, P. Szolovits, and H. C. Chueh, "Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach," *BMC Med. Inform. Decis. Mak.*, 2017, doi: 10.1186/s12911-017-0556-8.

[10] S. K. Yuwono, H. T. Ng, and K. Y. Ngiam, "Learning from the Experience of Doctors: Automated Diagnosis of Appendicitis Based on Clinical Notes," 2019, doi: 10.18653/v1/w19-5002.

[11] Y. Peng, S. Yan, and Z. Lu, "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets," no. iv, pp. 58–65, 2019, doi: 10.18653/v1/w19-5006.

[12] X. Dong *et al.*, "Deep learning for named entity recognition on Chinese electronic medical records: Combining deep transfer learning with multitask bi-directional LSTM RNN," *PLoS One*, vol. 14, no. 5, pp. 1–15, 2019, doi: 10.1371/journal.pone.0216046.

[13] I. Rish, "An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier," no. January 2001, pp. 41–46, 2014.

[14] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/bf00116251.

[15] T. T. Hastie, "The Elements of Statistical Learning Second Edition," *Math. Intell.*, 2017, doi: 111.

[16] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," *30th Int. Conf. Mach. Learn. ICML 2013*, no. PART 3, pp. 2347–2355, 2013.

[17] L. Yang, S. Hanneke, and J. Carbonell, "A theory of transfer learning with applications to active learning," *Mach. Learn.*, vol. 90, no. 2, pp. 161–189, 2013, doi: 10.1007/s10994-012-5310-y.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," no. Mlm, 2018, [Online]. Available: http://arxiv.org/abs/1810.04805.

[19] M. Peters *et al.*, "Deep Contextualized Word Representations," pp. 2227–2237, 2018, doi: 10.18653/v1/n18-1202.

[20] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 1, pp. 328–339, 2018, doi: 10.18653/v1/p18-1031.

[21] J. Liu, Z. Zhang, and N. Razavian, "Deep EHR: Chronic Disease Prediction Using Medical Notes," 2018, [Online]. Available: http://arxiv.org/abs/1808.04928.