Jan 2022

The Effect of Pre-trained Language Models on Critical Finding Detection from Radiology Reports

Gıyaseddin BAYRAK [a,1], Muhammed Şakir TOPRAK [b,c] Murat Can GANİZ [a]
Halife KODAZ [c] and Ural KOÇ [d]

[a] *Computer Engineering Department, Marmara University, Turkey*
[b] *Asset Management Information Systems Department, Ministry of Health, Turkey*
[c] *Computer Engineering Department, Konya Technical University, Turkey*
[d] *Ankara City Hospital, Turkey*

Abstract. Radiology reports has a significant role in disease diagnosis and management process. On some cases radiology report may indicate a critical finding on which the doctor attending the patient must take an immediate action. Detection of such cases from radiology reports automatically and thus very quickly using Artificial Intelligence, particularly Natural Language Processing, models is very important and may save lives. This is a novel study to detect critical findings in the context of brain hemorrhage detection on Turkish radiology reports. We use about 30.000 labeled Brain Hemorrhage Computed Tomography (CT) reports for the training of supervised models and about 190 thousand reports for the pre-training or fine-tuning of language models and embedding models. To the best of our knowledge, this is the first study to use large scale of Turkish radiology reports. Furthermore, we show the affect of fine-tuning on pre-trained language models and static embeddings on the performance, concluding that fine-tuning using domain specific data improve classification performance.

Keywords. Emergency finding detection, Language models, Brain Hemorrhage, Radiology, Natural Language Processing, Deep learning

1. Introduction

Due to the ever increasing workload of the health systems, physicians can spare less time on average to diagnose the patients. As a result there is a high demand for systems that support physicians. Radiology imaging plays an important role in the diagnosis and disease management. However, radiology requests disrupts the process as the doctor usually take care of other patients while waiting for the radiology results in the form of both the image and the radiology text report. Critical findings may not be handled quickly due to the high work load of the physician. In this study, we primarily tackle this problem by developing a deep learning based classifier to detect critical findings in radiology reports of the patients who require immediate attention of the physician in the context of brain

[1] Corresponding Author: Istanbul, Turkey; E-mail: giyaseddinalfarkh@marun.edu.tr.

hemorrhage. Essentially, we use reports of brain Computed Tomography (CT) scans, and we focus the reports of patients diagnosed with cerebral hemorrhage in the preliminary or final diagnosis in the TeleRadiology[2] system of the Turkish Ministry of Health. The task of prioritizing critical brain hemorrhage patients has been covered in the literature by studying different modalities of data. For instance, Ertuğrul et al. [1] and Guo et al. [2] use CT images for classifying different types of brain hemorrhage. Critical case detection from radiology reports is also studied, yet with different grounds. For example, Karthik et al. [3] survey studies on brain ischemic stroke detection using deep learning. The survey involves a comparison between 35 studies that obtain data with different imaging techniques. In another similar study for critical case detection, authors detect the survival probability on heart failures using the clinical BERT representations [4]. There are several studies which develops classification models for radiology reports. Kim et al. [5] introduce a multi-label classification model using a subset of MIMIC-III [6] dataset. They compare their method with several traditional machine learning and deep learning based classifiers such as LSTM, CNN. One interesting note from this study is that the human annotaters inter-agreement rates are not as high as believed, and therefore machine learning classifiers may actually exceed the performance of humans. Olthof et al. [7] provide a comparison of several deep learning classifiers with transformer based language model BERT,and conclude that BERT outperforms others on two datasets; Traumatic Fracture and Chest data.

There are relatively few studies on Turkish radiology reports. Although they worked medical publication abstracts instead of radiology reports, one interesting study by Çelikten et al. [8] provide a comparison between Multilingual BERT and Turkish BERT (BERTurk) models in the classification context and report better results for Turkish pre-trained BERT model. Bayrak et al. studied Epilepsy classification using bi-LSTM on small dataset of radiology reports from MRI. [9].

Our contributions can be summarized as follows: 1) An implementation for the critical non-traumatic hemorrhage detection from radiology reports. 2) A comparison between the baseline pre-trained FastText [10] and BERT [11] language models and task-specific fine-tuned variations of them. 3) One of the first studies to train deep learning models on large textual dataset consist of Turkish radiology reports.

2. Methods

In order to create our labeled dataset for supervised machine learning models, we focus brain CT reports labeled using 10_{th} version of the International Classification of Diseases (ICD-10) diagnostic codes. There are two ICD-10 codes for a report, one for preliminary diagnosis by the doctor before the diagnostic imaging and final diagnosis after examining the images and radiology report. For our dataset, reports obtained for the patients with a preliminary diagnosis of cerebral hemorrhage indicated with following ICD-10 codes; I60, I61, I62. From preliminary diagnosis codes we obtain about hundred thousand reports for brain hemorrhage. The reports whose preliminary and final diagnoses codes match are labeled as positive. Those are the critical cases since detecting brain hemorrhage requires instant attention of the attending physician to take immediate ac-

[2]Known as TeleTıp system https://teletip.saglik.gov.tr/

tion like transferring the patient to the intensive care unit, surgery, etc. After removing the duplicate and noisy reports, we have 15697 (Positive) and 21819 (Negative) reports. We use 20% of the labeled data as the test set and the rest is split between training and validation with the ratio 80% : 20%.

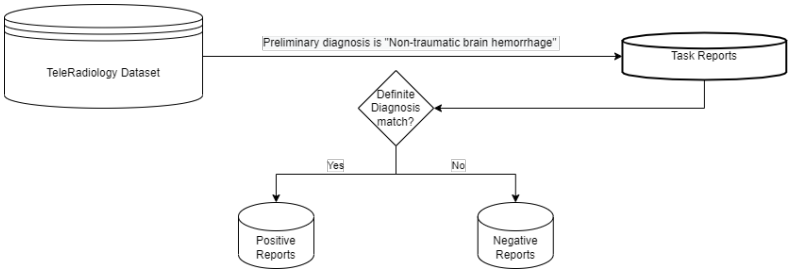

Figure 1. Formation procedure of the dataset.

Figure 1 shows the process of the labeled dataset formation for the training and evaluation of supervised machine learning based classification models.

For the unsupervised part, we use several neural language models for static and contextual word embeddings / dense distributed representations. For these models, we created a unlabeled dataset of about hundred ninety thousand radiological reports randomly chosen from brain and thorax CTs[3]. These are used for pre-training and fine-tuning of these unsupervised models.

One of our main purposes in this study to observe the affect of embedding models in the performance of deep learning models. Therefore we use the same classifier with the fixed hyper-parameters along the experiments, namely Bi-directional Long-Short Term Memory (BiLSTM) [12], [9], [13] and [14]. So all the other conditions, such as the data and hyper-parameters are constant, we change the embedding method and observe its affect on the performance of the classifier. A similar approach can be seen in [13]. A fully connected layer is constructed right after the embedding layer, and followed by an output layer of two neurons in which softmax is used [14]. A single BiLSTM layer, 0.5 recurrent dropout for BiLSTM layer, 0.4 dropout for the fully connected layer, 128 LSTM units, ADAM optimizer, learning rate 0.001 with no L2 penalty similar to [13], [9], [12]. All the classification trials was trained for 4 epochs as in .

As the embedding layer of the bi-LSTM we use following alternatives: 1) fastText pre-trained model for Turkish on common crawl [4] 2) fastText model trained from scratch using our unsupervised dataset of 190 thousand radiology reports 3) BERT pre-trained model for Turkish[5] 4) BERT pre-trained model for Turkish (as in 3) but fine-tuned as masked language model using our unsupervised dataset of about 190 thousand radiology reports 5) BERT pre-trained model for Turkish (as in 3) but fine-tuned as masked language model using training set portion of the supervised dataset of about 37 thousand reports (80% of this is used as training set). Furthermore, we also observe the affect

[3]Random sampling for thorax reports excluded COVID related ones indicated by U07.3, U07.1, U07.2 and U07.

[4]The publicly available model in the official fastText website https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.tr.300.bin.gz

[5]The publicly available model https://huggingface.co/dbmdz/bert-base-turkish-uncased

Pre-trained Embedding Layer	**Cased**	**Uncased**	**Freezed**
FastText common crawl	55.42	54.10	No
FastText common crawl	53.77	55.53	Yes
FastText pre-trained on 190k Rad unsupervised reports	54.14	55.02	No
FastText pre-trained on 190k Rad unsupervised reports	54.27	56.68	Yes
bi-LSTM Embeddings (without pre-trained embeddings)	65.45	68.06	No
BERT Base	68.03	69.17	Yes
BERT Fine-tuned on training reports	70.53	72.01	Yes
BERT Fine-tuned on 190k unsupervised reports	**73.31**	**72.27**	Yes

Table 1. F1 results of bi-LSTM classifier with different embedding layers for Turkish radiology report classification.

of freezing embedding layer for the fastText embeddings. We use F1 (a.k.a. F-measure) score as our main evaluation metric which is the harmonic mean of precision and recall. F1 scores are averaged among classes, that we are reporting the macro average F1.

3. Experiment Results

Our experiment results are mainly given in Table 1. We show the affect of using different embeddings in bi-LSTM in terms of F1 scores in this table. The second and third columns show the version of the training set. ”Cased” shows the results of the original dataset. On the other hand, ”Uncased” shows the results of the dataset that is converted to lowercase. Fourth column indicates if the embedding layer of the bi-LSTM is freezed or not. If it is freezed then the embedding layer values are not updated by the backpropagation algorithm.

4. Discussion

As can be seen from 1 the choice of embedding method has a drastic affect on the performance of the bi-LSTM. Among the fastText results we can see that pre-training with large amounts of domain specific data may yield a slightly better results. However, the fastText results with different parameters of cased and uncased data usage and freezing the embedding layer results in pretty close F1 values. When we look at the overall results, as expected, BERT contextual embeddings works better than static embeddings of fastText. BERT related experiments show that fine-tuning the general Turkish model with large amounts of unsupervised domain specific data makes a difference. The difference is most visible when we fine-tune BERT with smaller but labeled domain specific data. That is where we obtain the best F1 result of 73.31%.

5. Conclusion

We develop a novel deep learning based classification model critical finding detection in the context of brain hemorrhage diagnosis on large corpus of Turkish radiology reports. Furthermore, we analyze the affect of pre-trained static and contextual word embeddings on the performance of bi-LSTM classifier in this domain. Our experiments show the

fine-tuning of general pre-trained language model in Turkish with domain specific data improve the performance considerably. In the future, we plan to examine the impact of cross-language transfer learning on critical finding detection task. In addition, we aim to focus on maximizing the critical finding classification performance by hyper-parameter optimization, employing different architectures and deep learning models to develop a practical classification system. This, of course, may require an explainable system to verify the correctness of decisions of the model as in [4].

Acknowledgements

For this study, the use of data is approved under ethics vote number E.Kurul-E1-22-2326 by the ethical committee of the Ministry of Health. Authors thank Ministry of Health for providing data and computational resources. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the Ministry of Health of the Government of Turkey.

References

[1] Ertuğrul ÖF, Akıl MF. Detecting hemorrhage types and bounding box of hemorrhage by deep learning. Biomedical Signal Processing and Control. 2022;71:103085.

[2] Guo D, Wei H, Zhao P, Pan Y, Yang HY, Wang X, et al. Simultaneous classification and segmentation of intracranial hemorrhage using a fully convolutional neural network. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE; 2020. p. 118-21.

[3] Karthik R, Menaka R, Johnson A, Anand S. Neuroimaging and deep learning for brain stroke detection-A review of recent advancements and future prospects. Computer Methods and Programs in Biomedicine. 2020:105728.

[4] Lee HG, Sholle E, Beecy A, Al'Aref S, Peng Y. Leveraging Deep Representations of Radiology Reports in Survival Analysis for Predicting Heart Failure Patient Mortality. arXiv preprint arXiv:210501009. 2021.

[5] Kim BH, Ganapathi V. Read, Attend, and Code: Pushing the Limits of Medical Codes Prediction from Clinical Notes by Machines. In: Machine Learning for Healthcare Conference. PMLR; 2021. p. 196-208.

[6] Johnson A, Pollard T, Mark III R. MIMIC-III clinical database. Physio Net. 2016;10:C2XW26.

[7] Olthof A, van Ooijen P, Cornelissen L. Deep Learning-Based Natural Language Processing in Radiology: The Impact of Report Complexity, Disease Prevalence, Dataset Size, and Algorithm Type on Model Performance. Journal of medical systems. 2021;45(10):1-16.

[8] Çelıkten A, Bulut H. Turkish Medical Text Classification Using BERT. In: 2021 29th Signal Processing and Communications Applications Conference (SIU). IEEE; 2021. p. 1-4.

[9] Bayrak S, Yucel E, Takci H. Epilepsy Radiology Reports Classification Using Deep Learning Networks. COMPUTERS, MATERIALS AND CONTINUA : Tech Science Press. 2022;70(2):3589-607.

[10] Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 2017;5:135-46.

[11] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

[12] Graves A, Fernández S, Schmidhuber J. Bidirectional LSTM networks for improved phoneme classification and recognition. In: International conference on artificial neural networks. Springer; 2005. p. 799-804.

[13] Drozdov I, Forbes D, Szubert B, Hall M, Carlin C, Lowe DJ. Supervised and unsupervised language modelling in Chest X-Ray radiological reports. Plos one. 2020;15(3):e0229963.

[14] Bustos A, Pertusa A, Salinas JM, de la Iglesia-Vayá M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis. 2020;66:101797.