JOURNAL of
MATERIALS and MECHATRONICS:A

e-ISSN 2717-8811
JournalMM, 2025, 6(1), 1-14
https://doi.org/10.55546/jmm.1571384

Araştırma Makalesi / Research Article

Extraction of Clinical Entities from Chest Radiology Reports Using NLP Methods

Uçman ERGUN[1], Sedanur ORCIN[2*], Sezin BARIN[3]

[1] Afyon Kocatepe University, Faculty of Engineering, Department of Biomedical Engineering, Afyonkarahisar, Türkiye, ORCID ID: https://orcid.org/0000-0002-9218-2192, uergun@aku.edu.tr
[2*] Afyon Kocatepe University, Faculty of Engineering, Department of Biomedical Engineering, Afyonkarahisar, Türkiye, ORCID ID: https://orcid.org/0009-0007-4345-4984, orcinsedanur@gmail.com
[3] Afyon Kocatepe University, Faculty of Engineering, Department of Biomedical Engineering. Afyonkarahisar, Türkiye, ORCID ID: https://orcid.org/0000-0002-0394-2779, sbarin@aku.edu.tr

Geliş/ Received: 21.10.2024**;** **Revize/Revised:** 06.12.2024 **Kabul / Accepted:** 20.12.2024

ABSTRACT: Radiology reports are essential for clinical decision-making and diagnosis, containing complex and detailed information. However, their unstructured nature makes efficient processing and analysis challenging, increasing the workload of healthcare professionals and slowing down clinical workflows. Natural Language Processing (NLP) techniques provide effective solutions by extracting meaningful information from such texts, reducing expert workload, and expediting decision-making processes. This study focuses on Named Entity Recognition (NER) in chest radiology reports using the RadGraph dataset, annotated with four tag types. The objective is to compare the performance of two NLP models—BERT (Bidirectional Encoder Representations from Transformers) and LSTM (Long Short-Term Memory) —to identify the most suitable approach for clinical data. Various training parameters, including learning rate, optimization algorithm, and input size, were optimized to enhance model performance. To address the class imbalance in the dataset, data augmentation techniques were applied, and both models were fine-tuned. The results revealed that BERT, leveraging its attention mechanism, demonstrated superior performance in identifying complex terms and entities, outperforming LSTM in accuracy, precision, recall, and F1 score. While LSTM effectively captured long-term dependencies, it required longer training times. This research highlights the potential of NLP in automating the extraction of clinical entities from radiology reports. It provides valuable insights for optimizing models and developing clinical decision support systems, ultimately aiming to enhance the efficiency of healthcare workflows.

Keywords: Deep learning, Natural language processing, Named entity recognition, Radiological report, BERT

*Sorumlu yazar / Corresponding author: orcinsedanur@gmail.com
Bu makaleye atıf yapmak için /To cite this article

Ergün, U., Orcin, S., Barın S. (2024). Extraction of Clinical Entities from Chest Radiology Reports Using NLP Methods. Journal of Materials and Mechatronics: A (JournalMM), 6(1), 1-14.

1. INTRODUCTION

In recent years, artificial intelligence technologies have revolutionized many areas of life and attracted attention with innovative solutions. NLP technologies, one of the most prominent of these developments, are gaining more and more importance, especially in the healthcare sector. The complex expressions and findings often encountered in radiology reports can make it difficult to interpret the reports correctly, especially for inexperienced physicians, and can lead to misdiagnoses (Yamashita et al., 2022; Nishio et al., 2024). At this point, integrating NLP methods can play a critical role in overcoming these challenges. Integration of NLP into the healthcare system can provide significant improvements in diagnosis and treatment processes by offering various advantages to both physicians and patients.

This study focuses on the automatic, accurate and efficient labeling of important findings and comments in radiology reports using NLP techniques. The study examines how effective NER techniques in these reports are and how these techniques perform with different models. The hypothesis of the study is that by integrating BERT and LSTM models with NER techniques, critical information in radiology reports can be accurately and quickly labeled and the most effective method can be determined by comparing the performance of these models.

As a result of the literature review, the RadGraph dataset was studied (RadGraph Dataset, 2021). When the dataset was analyzed, it was seen that there was an unbalanced distribution between the number of tags ("ANAT-DP=5366", "OBS-DP=5046", "OBS-U=584", and "OBS-DA=1389"). To overcome this problem, the dataset was expanded. In the data preprocessing stage, the data were processed according to the needs of the selected dataset in order to present the data more consistently and to improve the performance of the model (M. Wang and Hu, 2021; Uskaner Hepsağ et al., 2023). After 2019, most of the NLP studies focus on deep learning architectures (Nag et al., 2024; C. Pereira et al., 2024). In this context, both RNN (Recurrent Neural Network) based LSTM model and Transformer architecture-based BERT model are considered in our study (Uskaner Hepsağ et al., 2023; Rani et al., 2024). Different training parameters were tested on the models and the most successful performing parameters were determined (Yan et al., 2022).

As a result, when this study was completed, significant progress was made in the automatic analysis of radiology reports and the extraction of important findings. By comparing BERT and LSTM models, it became possible to determine the most appropriate NLP methods for supporting the diagnosis and treatment processes of physicians.

2. MATERIALS AND METHODS

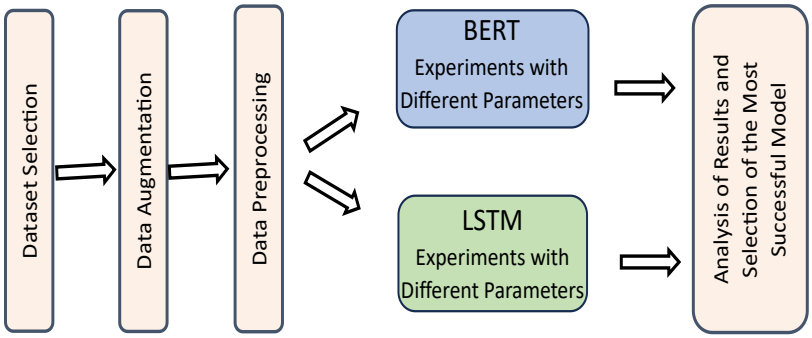


Figure 1. Workflow applied in the study

The study focuses on NER, which aims to speed up the evaluation of radiology reports by physicians and reduce errors in diagnoses. The steps planned in the study process were implemented as shown in the workflow diagram in Figure 1.

2.1 Dataset Selection

The study was conducted with the Radgraph (RadGraph Dataset, 2021) dataset, which includes the MIMIC-CXR and CheXpert datasets offered in the PhysioNET Database. In order to access the dataset, it was requested to complete the trainings prepared by the dataset providers and to achieve a 90% success rate in the exams. This process consisting of 2 sections and 16 modules was completed and the dataset was accessed. The RadGraph dataset was created by tagging chest radiology reports from the MIMIC-CXR and CheXpert datasets. The MIMIC-CXR dataset was created in collaboration with Massachusetts General Hospital and MIT Laboratory for Computational Physiology and de-identified personal health information (PHI) in accordance with HIPAA requirements. The CheXpert dataset was developed by Stanford University and was similarly de-identified in accordance with HIPAA requirements, and PHI was replaced with fake PHI using automated and manual methods. During the development of the RadGraph dataset, the MIMIC-CXR and CheXpert datasets were used to identify radiology reports entity names and relationships. In the tagging process, three radiologists tagged the reports according to the schema developed by Dr. Curt Langlotz on the Datasaur.ai platform. As a training set, 425 MIMIC-CXR reports were used, 75 reports were used for development and 50 MIMIC-CXR and 50 CheXpert reports were used for testing. The dataset used includes four entities—ANAT-DP (Anatomical Descriptor Present), OBS-DP (Observation Descriptor Present), OBS-DA (Observation Descriptor Absent), and OBS-U (Observation Descriptor Uncertain)—as well as three relationship types, aimed at structuring clinical information in radiology reports. Four entity labels were used in the study.

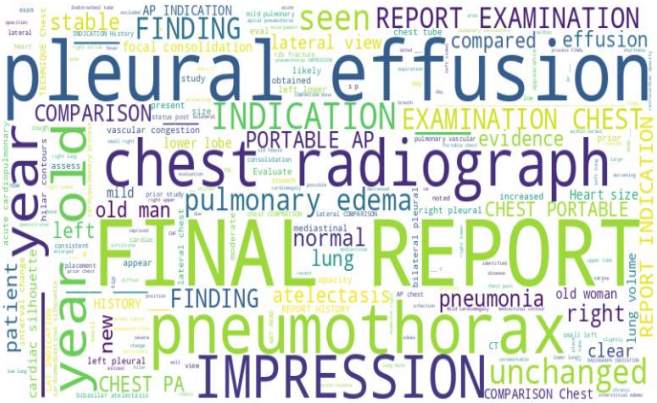


Figure 2. Commonly used words from the RadGraph dataset

2.2 Model Selection and Parameter Settings

The analysis of radiological reports has become the focus of deep learning methods today. The literature in this field reveals that RNN and Transformer deep learning architectures are increasingly used for processing radiology reports (Sun et al., 2023). These methods are of great importance for understanding the complexity of text data, extracting the information they contain, and effectively classifying reports. Deep learning techniques offer powerful tools for obtaining valuable insights from radiology reports (C. Pereira et al., 2024).

RNN is a deep learning model for processing sequential data. Especially when used in areas such as time series data and natural language processing, it takes into account the sequential structure of the data and associates the information from previous steps with the current steps. Because of this feature, it provides successful results in language processing problems (Zhang et al., 2018).

The "Transformer" architecture is a deep learning model based on the attention mechanism developed by Google in 2017 and presented in the paper "All You Need is Attention" (Vaswani et al., 2017). Traditionally used sequential processing structures such as RNN or Long Short-Term Memory (LSTM) are replaced by attention-based mechanisms in the Transformer architecture. Attention-based mechanisms are particularly notable for their "Multi-Head Attention" structure. This structure ensures that attention-oriented vectors are generated for each input token. The input tokens and vectors are combined to create an output. The importance of each token in the text relative to other tokens is determined and the contextual relations of the text are modeled more effectively (Brasoveanu & Andonie, 2020). The Transformer model has revolutionized the field of natural language processing and has achieved the best results in many tasks (Rahali and Akhloufi, 2023).

This study focuses on deep learning architectures such as the LSTM model based on RNN architecture and the BERT model based on "Transformer" architecture. The features of the models used in the study are given below.

2.2.1 LSTM (Long short-term memory)

LSTM is a special kind of RNN family and is designed to solve the problem of long-term dependency over time. It works using specialized memory units called cells. These cells can control information through input, output and forget gates. Thanks to these structures, LSTM can more effectively learn long-term dependencies and relationships in extensive texts (M. Tarwani and Edem, 2017). In this study, the performance of the LSTM model will be evaluated using an English-language dataset in the medical field. The model scheme presented in Figure 3 forms the basis of the study and the input representation is integrated into the architecture of the model (Uskaner Hepsağ et al., 2023).

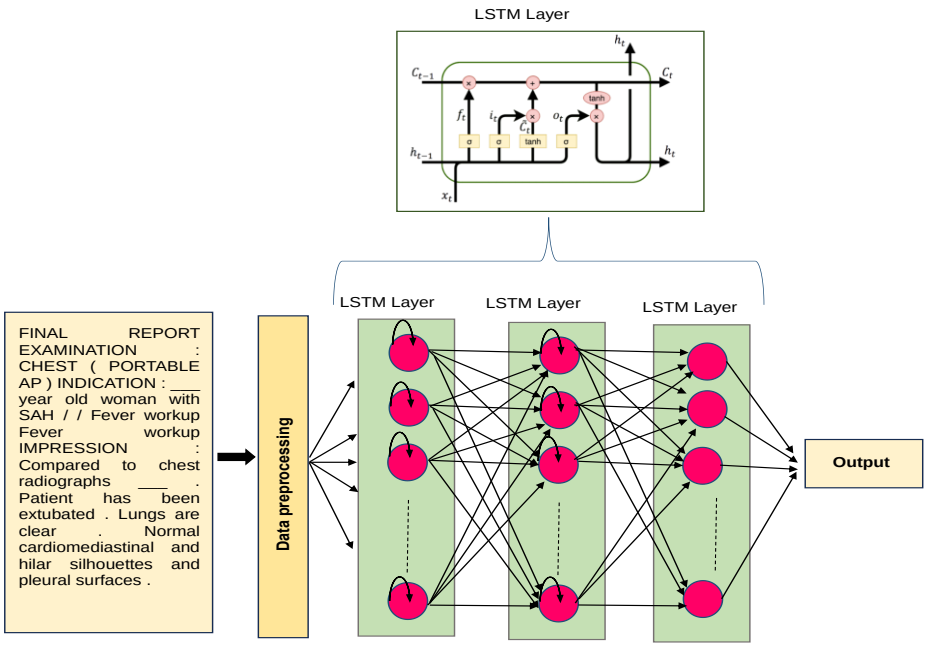


Figure 3. Schematic of the LSTM model used in the study

At the core of the LSTM architecture are the “Cell State” and various gates. “Cell State” is the channel through which information is carried in memory. The gates decide which information is important. The Forget Gate makes decisions about previous knowledge (ht) and current knowledge (xt). Sigmoid activation determines which information to forget. Input Gate updates the cell state. Sigmoid operation decides which information to update. Tanh activation organizes the data. The information to be updated is determined by multiplying the results. Output Gate determines the input of the next cell (ht+1). It is also used for prediction. The sigmoid operation helps to determine the input, while the tanh operation determines the state of the current information. Using these mechanisms, LSTM can effectively learn long-term dependencies and process complex language structure. Therefore, it is frequently used in the field of natural language processing, especially in tasks such as text prediction, translation, sentiment analysis, Entity Name Extraction (Rahman et al., 2021).

2.2.2 BERT (Bidirectional encoder representations from transformers)

BERT is a deep learning model based on the Transformer architecture developed by Google, which is an important milestone in the field of natural language processing (Vaswani et al., 2017). It is based on the masked language model. Some words in the text are subjected to random masking to improve the model's ability to understand the context. Its main architecture is a bidirectional transformer encoder. When processing a text by a given language model, allows to learn the context of each word with the influence of both preceding and following words. It provides a more powerful model that can use a wider context to determine the meaning of a word in a text. BERT is pre-trained on a large training dataset. In the training phase, a large amount of text data is used to improve the model's overall language understanding capability (Turchin et al., 2023). Figure 4 shows the schematic of the BERT model used in this study (Uskaner Hepsağ et al., 2023).

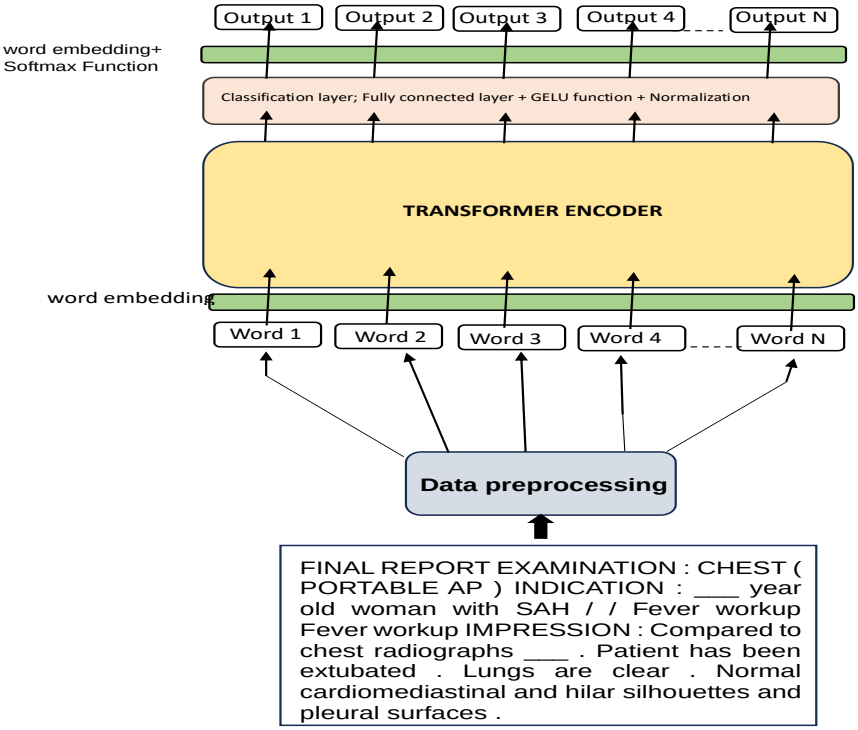


Figure 4. Schematic of the BERT model used in the study

In the diagram, the input layer receives the text of the radiological report and creates a basic structure to operate on this text. The word embedding process converts the words into numeric values to better capture the semantic relations of the text and enable the model to process the words in a more meaningful way. The encoder block processes the words in the text and extracts their features. The parallel attention mechanism determines the importance of each word and feature, allowing the model to give more weight to important words. In this way, the model can focus more on critical information in the text for more effective entity name extraction.

2.3 Data Augmentation

In this study, a BERT-based data augmentation technique was applied to increase the representation of classes with low tag counts (Abuzayed et al., 2021; Liu et al., 2022). To address the tag imbalances in the dataset, several steps were taken to reach the target number of tags for each class. First, the amount of boosting required to reach the targeted number was calculated by considering the current number of each tag. As presented in Table 1, sentences with a small number of tags in the raw dataset were extracted from the dataset and certain words in these sentences were masked. This was done by preserving the context of the sentence. The BERT model was used to predict the masked words. The BERT model performed possible word predictions based on the context of the masked word and the most likely word predictions were selected. These predicted words were used to replace the missing words in the masked sentences and new sentences were generated. Finally, the new sentences generated in this way and the corresponding tags were added to the dataset. As a result of this process, the number of tags in the augmented dataset presented in Table 1 was reached. Because of this approach, the number of labels with a small number of tags increased, allowing the model to learn them better.

The use of the BERT model played a critical role in gaining a deeper understanding of the language context and making accurate word predictions. This data augmentation method can be considered as an effective strategy to improve the performance of the model in class imbalanced datasets, especially in the field of natural language processing (NLP).

Table 1. Raw- Augmented dataset tag counts

Tag	Raw Data Set	Augmented Data Set
ANAT-DP	5366	5366
OBS-DP	5046	5046
OBS-DA	1389	6041
OBS-U	584	3904

2.3 Evaluation Metrics

The performances of the LSTM and BERT-based models used in this study are compared through the evaluation metrics presented in Table 2.

Table 2. Evaluation Metrics

Metric	Formula	Description
Accuracy	Accuracy=Correct Predictions/ Total Predictions	Represents the proportion of correctly identified entities out of the total extracted entities. This metric measures the overall success of the model.
Precision	Precision= True Positives / True Positives + False Positives	Indicates how many of the entities identified by the model are correct. High precision reflects the model's ability to make accurate predictions.

Table 2. Evaluation Metrics (continued)

Metric	Formula	Description
Recall	Recall= True Positives/ True Positives + False Negatives	Shows how well the model can identify actual entities present in the reports. High recall indicates that the model is capable of capturing most of the true entities.
F1 Score	F1 Score=2×Precision+Recall/ Precision × Recall	The harmonic means of precision and recall, the F1 score balances these two metrics, particularly in cases where the dataset has imbalanced labels. This metric assesses whether the model performs well in both precision and recall.

3. RESULTS AND DISCUSSION

3.1 Results

This study was performed on the Google Colab Pro+ platform using the A100 GPU. Various hyperparameter combinations were examined on LSTM and BERT models. The parameters that provide the highest performance were optimized. With these parameters, data augmentation methods were applied and model training and testing were performed on the enriched data set obtained at the end of this process. The experimental studies aimed to maximize the performance of the models and provide the highest accuracy rate on medical data.

3.1.1 Determination of the optimization algorithm

In the first stage, the performance of ADAM (Adaptive Moment Estimation) and SGD (Stochastic Gradient Descent) optimization algorithms were evaluated in detail using LSTM and BERT models on the RadGraph dataset. In the training process, early stopping technique was applied and training was performed for a total of 50 epochs. The impact of both algorithms is compared in terms of performance metrics such as model accuracy, loss and processing time. The test results and the success levels of the optimization algorithms are presented in Table 3.

Table 3. Performance results of different optimization algorithms in BERT and LSTM models

	BERT		**LSTM**	
	ADAM	SGD	ADAM	SGD
Accuracy (%)	89.25	90.10	78.21	79
Precision (%)	87.10	87.96	75.36	76.21
Recall (%)	88.40	88.75	76.45	77.14
F1 Score (%)	87.75	88.35	75.90	76.67

As a result of the analysis of the test results, no significant performance difference was found between the ADAM and SGD optimization algorithms. At the same time, the SGD algorithm resulted in longer training times. Therefore, in order to minimize the computational cost, the studies were continued with the ADAM optimization algorithm.

3.1.2 Determining the learning rate

Within the scope of the study, the learning rates presented in Table 4 were tested with the ADAM optimization algorithm. As a result of the experiments, the learning rate providing the highest performance was determined and presented in Table 5.

Table 4. Learning rates used in the study

Literature	Learning Rate
Houlsby et al., 2019	1×10^{-4}
Lamproudis et al.,2021	1×10^{-5}
Choi et al., 2020	2×10^{-5}

As a result of the examination of the test results, it was determined that the learning rate providing the highest success was 1×10^{-5} in the BERT model and 1×10^{-4} in the LSTM model. In line with these findings, the studies were continued on the learning rates that provided the highest success.

Table 5. Performance results of different learning rates in BERT and LSTM models.

	BERT			**LSTM**		
	2×10^{-5}	**1×10^{-5}**	**1×10^{-4}**	**2×10^{-5}**	**1×10^{-5}**	**1×10^{-4}**
Accuracy (%)	89.25	90.24	90.10	78.21	79.56	79.74
Precision (%)	87.10	89.34	88.80	75.36	78.12	78.99
Recall (%)	88.40	88.98	88.98	76.45	78.96	79.90
F1 Score (%)	87.75	89.16	88.89	75.90	78.54	79.44

3.1.3 Determining the learning rate

The effects of different input sizes (64, 128, 256) on the performance of the models were investigated. In the training process, the optimization algorithm, learning rate and epoch parameters determined in the previous steps were used. The results obtained are presented in Table 6 to evaluate the effects of input sizes on model performance.

Table 6. Effect of different input sizes on model performance in BERT and LSTM models

	BERT			**LSTM**		
	64	**128**	**256**	**64**	**128**	**256**
Accuracy (%)	85.69	90.24	90.84	76.52	79.74	80.11
Precision (%)	89.42	89.34	89.97	77.75	78.99	79.45
Recall (%)	85.63	88.98	90.26	76.69	79.90	79.68
F1 Score (%)	87.48	89.16	90.11	77.22	79.44	79.56

3.1.4 Model performance evaluation on augmented data set

As a result of the training performed with the RadGraph dataset using the parameter values specified in the previous work packages, the optimum hyperparameters were determined. In line with these optimum parameters, the model training was performed on the new data set created by the data augmentation process applied to the RadGraph data set. The training results are presented in Table 7.

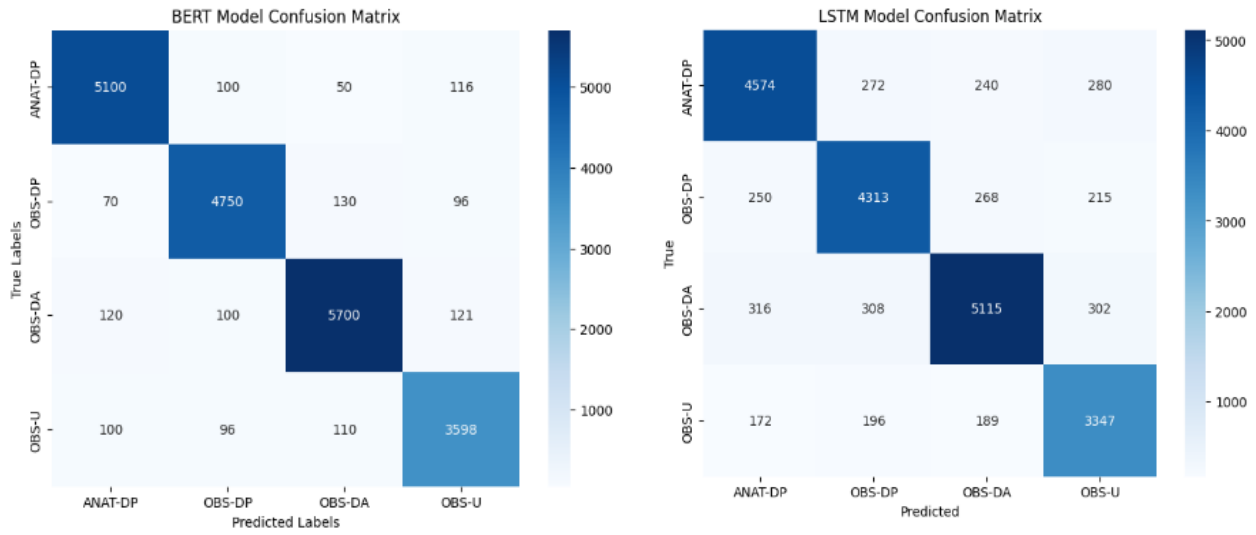
Table 7 Performance results of LSTM and BERT models on the augmented dataset

	BERT	**LSTM**
Accuracy (%)	95.48	85.26
Precision (%)	94.23	83.57
Recall (%)	96.69	84.88
F1 Score (%)	95.44	84.22

According to the training results, as observed in the RadGraph dataset, the BERT model achieved a higher success rate compared to the LSTM model. While the F1 score of the BERT model was 95.44, the F1 score of the LSTM model was 84.22.

Figure 5(a) and (b) show the confusion matrices of the LSTM and BERT models, respectively. When the confusion matrices are analyzed, it is seen that the BERT model misclassifies medical entities such as ANAT-DP, OBS-DP, OBS-DA, and OBS-U much less than the LSTM model. It is observed that the BERT model recognizes common entities such as ANAT-DP and OBS-DP with high accuracy. On the other hand, the LSTM model has higher error rates, especially in the rarer OBS-DA and OBS-U classes. This is evident in the LSTM confusion matrix in Figure 5, where the false positive and false negative rates are more pronounced.

One of the main reasons why the BERT model is more successful is its transformer-based structure. BERT is better able to model long-distance dependencies between words in the text, and thanks to its bidirectional language model, it produces more accurate results by taking into account both the preceding and following context. Especially in datasets where contextual information is critical, such as medical reports, the high performance of the BERT model makes it more advantageous than the LSTM.

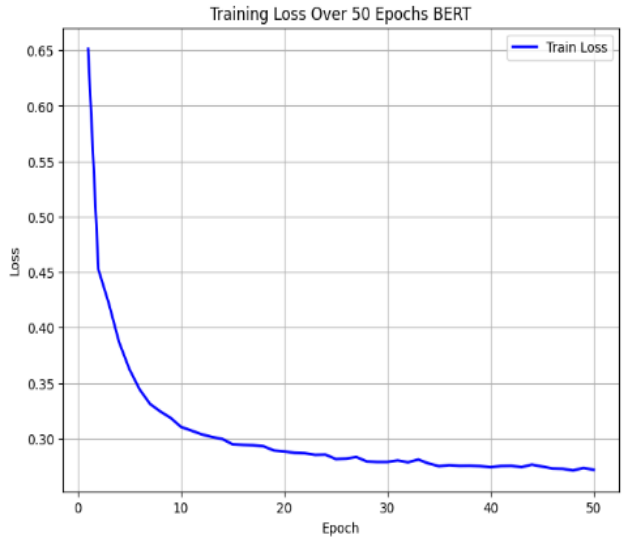


(a) (b)

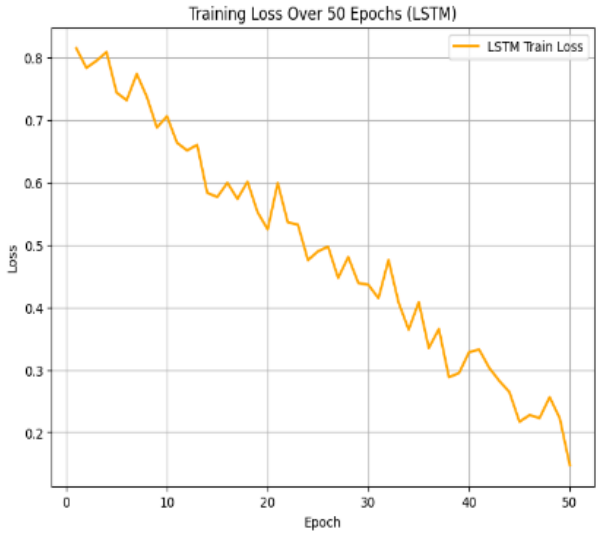
Figure 5. (a) Confusion matrix of BERT model, (b) Confusion matrix of LSTM model

Figures 6(a) and (b) show the loss function plots of the LSTM and BERT models respectively during the training process. It is observed that the loss curve of the LSTM model is wavy and slow. This wavy structure indicates that the LSTM model experiences instability in the learning process and has difficulty in optimizing the parameters of the model. The sequential processing structure of the LSTM can cause gradient loss in long sentences and complex contextual relations. This slows down the learning process of the model, especially for long and complex texts such as the one used in this dataset, and causes it to face more noise. In contrast, the loss curve of the BERT model was smoother and faster. Due to the advantages of the attention mechanisms, the BERT model learned contextual information more effectively at each step and reduced errors faster. This steady reduction suggests that the model generalizes the dataset better and produces more optimized results at each

learning step. Moreover, this regular structure in BERT's training process enabled it to achieve higher performance in less time and significantly reduce training loss.



(a)



(b)

In this study, while comparing the performance of BERT and LSTM models on the augmented RadGraph dataset, the basic metrics of precision, recall, accuracy and F1 score are taken into account. The results show that the BERT model outperforms the LSTM model in all of these metrics. The higher precision of the BERT model indicates that the model increases the number of true positive classifications and decreases the number of false positive predictions. This is particularly evident for the ANAT-DP and OBS-DP tags. The fact that the BERT model classifies entities in medical reports more carefully and accurately provides a significant advantage in preventing misdiagnosis. In terms of recall, the BERT model was also more successful than the LSTM. The BERT model successfully identified a large proportion of positive examples that should be labeled correctly. Especially in medical texts, the recall rate is critical to prevent false negatives. Since false negatives can have serious consequences, for example if a disease is missed, this superiority in recall rate shows that the BERT model offers a significant advantage for medical data analysis.

Finally, the F1 score measures the overall performance of the model by balancing both metrics, precision and recall. The BERT model has a high F1 score of 95.44%, indicating that the overall performance of the model is consistent and reliable. These results show that the BERT model is more effective than the LSTM in the task of named entity recognition in medical texts and provides more accurate results with fewer errors.

3.2 Discussion

The studies conducted in literature were examined and compared with the developed model as shown in Table 8.

Table 8. Comparison of this study with other studies in the literature

Study	Dataset	Dataset Privacy	Method	Results (F1 score)
Thurkal et al. (2023)	Chest X-Ray	Private	BERT	78.97
Jain et al. (2021)	RadGraph	Public	PubMed BERT	0.86

Table 8. Comparison of this study with other studies in the literatüre (continued)

Study	Dataset	Dataset Privacy	Method	Results (F1 score)
López-Úbeda et al. (2020)	Chest CT	Private	LSTM	75.77
Yuan et al. (2019)	Chest X-Ray	Public	CNN and LSTM	CNN 0.90 LSTM 0.90
Banerjee et al. (2019)	Chest CT	Private	RNN	0.77
Cornegruta et al. (2016)	Chest X-Ray	Private	BiLSTM	0.90
Proposed Method	Augmented RadGraph	Public	LSTM and BERT	LSTM 0.84 BERT 0.95

When the table presented above is examined, it is seen that Thurkal et al. on a special dataset of chest X-ray radiology reports, an F1 score of 78.97 was reported using the BERT model (Thurkal et al., 2023). In another study with the RadGraph dataset, Jain et al. utilized the PubMed BERT model and obtained an F1 score of 0.86 (Jain et al., 2021). López-Úbeda et al. analyzed chest CT radiology reports with the LSTM model and achieved an F1 score of 75.77 (López-Úbeda et al., 2020). In the study by Yuan et al. in 2019, both CNN and LSTM models were tested on Chest X-Ray reports and an F1 score of 0.90 was recorded in both models (Yuan et al., 2019). Banerjee et al. reported an F1 score of 0.77 in their study with the RNN model on Chest CT reports (Banerjee et al., 2019). In 2019, Cornegruta et al. obtained a successful result with an F1 score of 0.90 using BiLSTM model on private Chest X-Ray radiology reports (Cornegruta et al., 2016).

In this study, unlike other studies in the literature, a comprehensive parameter optimization is performed on both BERT and LSTM models. These parameters are applied to the publicly available RadGraph dataset. The unbalanced label distribution in the RadGraph dataset was balanced with data augmentation techniques to improve the performance of the model. A comparison was made between the BERT and LSTM models using this newly created balanced dataset, and the results showed that the BERT model not only outperformed the LSTM model with a 95% success rate, but also outperformed other studies in the literature (Tokgoz et al.,2021; Yang et al.,2019).

4. CONCLUSIONS

In this study, BERT and LSTM models are studied using the publicly available RadGraph dataset of chest radiology reports. In the first stage, different optimization algorithms were tested for both models and it was determined that the ADAM algorithm gave the best results for both models. Then, various experiments were performed on the learning rate parameter and optimization was performed to determine the optimal value. In addition, as a result of the comparisons made on the maximum length parameter, a value of 256 was selected as the most appropriate parameter. After determining the parameters, improvements were made to the dataset in order to minimize the label imbalance in the dataset. The data augmentation process enabled the model to learn rare classes better and thus increased the overall performance rate. At this stage, labels with a small number of instances were increased to homogenize the overall distribution in the dataset. The regularized dataset was tested in BERT and LSTM models in line with the specified parameters. The comparison results showed that the BERT model performed better than the LSTM model, with an F1 score of 95% for the BERT model and 84% for the LSTM model. In the future, this study is planned to be optimized with different parameters to further improve the F1 score. In addition, the results obtained have the

potential to be integrated with Hospital Information Management Systems (HIMS) to provide support to specialist physicians.

5. ACKNOWLEDGEMENTS

This study was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) within the scope of the 2210-C National Graduate Scholarship Program.

6. CONFLICT OF INTEREST

Authors approve that to the best of their knowledge, there is not any conflict of interest or common interest with an institution/organization or a person that may affect the review process of the paper.

7. AUTHOR CONTRIBUTION

Uçman ERGÜN contributed to the Determining the concept and/or design process of the research. Sezin BARIN and Uçman ERGÜN contributed to the Management of the concept and/or design process of the research. Sedanur ORCİN contributed to the Data Collection. Sedanur ORCİN and Sezin BARIN contributed to the Data analysis and interpretation of the results. Uçman ERGÜN and Sedanur ORCİN contributed to the Preparation of the manuscript. Sezin BARIN contributed to the Critical analysis of the intellectual content. Uçman ERGÜN, Sedanur ORCİN, and Sezin BARIN contributed to the Final approval and full responsibility.

8. REFERENCES

Abuzayed A., Al-Khalifa H., Sarcasm and sentiment detection in Arabic tweets using BERT-based models and data augmentation. In Proceedings of the sixth Arabic natural language processing workshop 312-317, 2021.

Banerjee I., Ling Y., Chen M. C., Hasan S. A., Langlotz C. P., Moradzadeh N., Chapman B., Amrhein T., Mong D., Rubin D. L., Farri O., Lungren M. P., Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. Artificial Intelligence in Medicine 97, 79–88, 2019. https://doi.org/10.1016/j.artmed.2018.11.004

Brasoveanu A. M. P., Andonie R., Visualizing Transformers for NLP: A Brief Survey, 24th International Conference Information Visualisation (IV), Melbourne/Australia, September 07-11, 2020, pp: 270–279. https://doi.org/10.1109/IV51561.2020.00051

Choi H., Kim J., Joe S., Gwon Y., Evaluation of bert and albert sentence embedding performance on downstream nlp tasks, In 2020 25th International conference on pattern recognition (ICPR), Milan/Italy, January 10-15, 2021, pp: 5482-5487. 10.1109/ICPR48806.2021.9412102

Cornegruta S., Bakewell R., Withey S., Montana G., Modelling radiological language with bidirectional long short-term memory networks. arXiv preprint arXiv:1609.08409, 2016.

Houlsby N., Giurgiu A., Jastrzebski S., Morrone B., De Laroussilhe Q., Gesmundo A., Gelly S., Parameter-efficient transfer learning for NLP. 36th International Conference on Machine Learning, Long Beach/California, 2019, pp: 2790-2799. https://doi.org/10.1007/978-3-030-77211-6_12

Jain S., Agrawal A., Saporta A., Truon S. Q., Duong D. N., Bui T., Rajpurkar P., Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463, 2021.

Lamproudis A., Henriksson A., Dalianis H., Developing a clinical language model for Swedish: continued pretraining of generic BERT with in-domain data, In International Conference Recent Advances in Natural Language Processing (RANLP'21), Shoumen, September 1-3, 2021, pp: 790-797, 2021.

Liu J., Chen Y., Xu J., Low-Resource NER by Data Augmentation with Prompting, Thirty-First International Joint Conference on Artificial Intelligence, July 23-29, 2022, pp: 4252-4258.

López-Úbeda P., Díaz-Galiano M. C., Martín-Noguerol T., Luna A., Ureña-López L. A., Martín-Valdivia M. T., COVID-19 detection in radiological text reports integrating entity recognition. Computers in Biology and Medicine 127, 104066, 2020. https://doi.org/10.1016/j.compbiomed.2020.104066

López-Úbeda P., Martín-Noguerol T., Luna A., Automatic classification and prioritisation of actionable BI-RADS categories using natural language processing models. Clinical Radiology 79(1), e1-e7, 2024. https://doi.org/10.1016/j.crad.2023.09.009

Nag P. K., Bhagat A., Priya R. V., Khare D. kumar. Emotional Intelligence Through Artificial Intelligence: NLP and Deep Learning in the Analysis of Healthcare Texts, arXiv preprint arXiv: 2403.09762, 2024. http://arxiv.org/abs/2403.09762

Nishio M., Matsunaga T., Matsuo H., Nogami M., Kurata Y., Fujimoto K., Sugiyama O., Akashi T., Aoki S., Murakami T., Fully automatic summarization of radiology reports using natural language processing with large language models. Informatics in Medicine Unlocked 46, 101465, 2024. https://doi.org/10.1016/j.imu.2024.101465

Pereira S. C., Mendonça A. M., Campilho A., Sousa P., Lopes C. T., Automated image label extraction from radiology reports—A review. Artificial Intelligence in Medicine 149, 102814, 2024. https://doi.org/10.1016/j.artmed.2024.102814

RadGraph Dataset. Last Access Date: 13 Haziran 2024 from https://physionet.org/content/radgraph/1.0.0/

Rahali A., Akhloufi M. A., End-to-End Transformer-Based Models in Textual-Based NLP. AI, 4(1), 54–110, 2023. https://doi.org/10.3390/ai4010004

Rahman M. H., Islam M. S., Jowel M. M. U., Hasan M. M., Latif S., Classification of Book Review Sentiment in Bangla Language Using NLP, Machine Learning and LSTM, 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur/India, July 06-08, 2021, IEEE- 51525. https://doi.org/10.1109/ICCCNT51525.2021.9580116

Rani S., Jain A., Kumar A., Yang G., CCheXR-Attention: Clinical concept extraction and chest x-ray reports classification using modified Mogrifier and bidirectional LSTM with multihead attention. International Journal of Imaging Systems and Technology, 34(1), 1-15, 2024. https://doi.org/10.1002/ima.23025

Sun Z., Lin M., Zhu Q., Xie Q., Wang F., Lu Z., Peng Y., A scoping review on multimodal deeplearning in biomedical images and texts. Journal of Biomedical Informatics 146, 104482, 2023. https://doi.org/10.1016/j.jbi.2023.104482

Tarwani K. M., Edem S., Survey on Recurrent Neural Network in Natural Language Processing. International Journal of Engineering Trends and Technology 48(6), 301-304, 2017. https://doi.org/10.14445/22315381/IJETT-V48P253

Thukral A., Dhiman S., Meher R., Bedi P., Knowledge graph enrichment from clinical narratives using NLP, NER, and biomedical ontologies for healthcare applications. International Journal of Information Technology, 15(1), 53-65, 2023.

Tokgoz M., Turhan F., Bolucu N., Can B., Tuning language representation models for classification of Turkish news, 2021 International symposium on electrical, electronics and information engineering, 2021, pp: 402-407.

Turchin A., Masharsky S., Zitnik M., Comparison of BERT implementations for natural language processing of narrative medical documents. Informatics in Medicine Unlocked 36, 101139, 2023. https://doi.org/10.1016/j.imu.2022.101139

Uskaner Hepsağ P., Özel S. A., Dalcı K., Yazıcı A., Using BERT models for breast cancer diagnosis from Turkish radiology reports. Language Resources and Evaluation, 58, 981-1012 2024. https://doi.org/10.1007/s10579-023-09669-w

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention Is All You Need, arXiv preprint arXiv: 1706.03762, 2017. http://arxiv.org/abs/1706.03762

Wang M., Hu F., The application of nltk library for python natural language processing in corpus research. Theory and Practice in Language Studies 11(9), 1041-1049, 2021. https://doi.org/10.17507/tpls.1109.09

Yamashita R., Bird K., Cheung P. Y. C., Decker J. H., Flory M. N., Goff D., Morimoto L. N., Shon A., Wentland A. L., Rubin D. L., Desser T. S., Automated Identification and Measurement Extraction of Pancreatic Cystic Lesions from Free-Text Radiology Reports Using Natural Language Processing. Radiology: Artificial Intelligence 4(2), e210092, 2022.

Yan A., McAuley J., Lu X., Du J., Chang E. Y., Gentili A., Hsu C. N., RadBERT: Adapting Transformer-based Language Models to Radiology. Radiology: Artificial Intelligence 4(4), e210258, 2022. https://doi.org/10.1148/ryai.210258

Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R. R., Le Q. V., Xlnet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, 32, 10, 2019.

Yuan J., Liao H., Luo R., Luo J., Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11769 LNCS, 721-729, 2019. https://doi.org/10.1007/978-3-030-32226-7_80

Zhang X., Chen M. H., Qin Y., NLP-QA Framework Based on LSTM-RNN, 2nd International Conference on Data Science and Business Analytics (ICDSBA), Changsha/China, September 21-23, 2018, 307-311, 2018. https://doi.org/10.1109/ICDSBA.2018.00065