Diagn Interv Radiol 2025; DOI: 10.4274/dir.2025.243100

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Deep learning for named entity recognition in Turkish radiology reports

Abubakar Ahmad Abdullahi[1*]
Murat Can Ganiz[1*]
Ural Koç[2*]
Muhammet Batuhan Gökhan[2**]
Ceren Aydın[2**]
Ali Bahadır Özdemir[2**]

[1]Marmara University Faculty of Engineering, Department of Computer Engineering, İstanbul, Türkiye

[2]Ankara Bilkent City Hospital, Clinic of Radiology, Ankara, Türkiye

PURPOSE

The primary objective of this research is to enhance the accuracy and efficiency of information extraction from radiology reports. In addressing this objective, the study aims to develop and evaluate a deep learning framework for named entity recognition (NER).

METHODS

We used a synthetic dataset of 1,056 Turkish radiology reports created and labeled by the radiologists in our research team. Due to privacy concerns, actual patient data could not be used; however, the synthetic reports closely mimic genuine reports in structure and content. We employed the four-stage DYGIE++ model for the experiments. First, we performed token encoding using four bidirectional encoder representations from transformers (BERT) models: BERTurk, BioBERTurk, PubMedBERT, and XLM-RoBERTa. Second, we introduced adaptive span enumeration, considering the word count of a sentence in Turkish. Third, we adopted span graph propagation to generate a multidirectional graph crucial for coreference resolution. Finally, we used a two-layered feed-forward neural network to classify the named entity.

RESULTS

The experiments conducted on the labeled dataset showcase the approach's effectiveness. The study achieved an F1 score of 80.1 for the NER task, with the BioBERTurk model, which is pre-trained on Turkish Wikipedia, radiology reports, and biomedical texts, proving to be the most effective of the four BERT models used in the experiment.

CONCLUSION

We show how different dataset labels affect the model's performance. The results demonstrate the model's ability to handle the intricacies of Turkish radiology reports, providing a detailed analysis of precision, recall, and F1 scores for each label. Additionally, this study compares its findings with related research in other languages.

CLINICAL SIGNIFICANCE

Our approach provides clinicians with more precise and comprehensive insights to improve patient care by extracting relevant information from radiology reports. This innovation in information extraction streamlines the diagnostic process and helps expedite patient treatment decisions.

KEYWORDS

Named entity recognition, radiology reports, bidirectional encoder representations from transformers, Turkish, computed tomography, thorax

***Joint first authors**

****Contributed equally to this work**

Corresponding author: Ural Koç

E-mail: dr_uralkoc@hotmail.com

Received 30 October 2024; revision requested 10 December 2024; last revision received 01 January 2025; accepted 13 January 2025.

Epub: 28.02.2025

DOI: 10.4274/dir.2025.243100

Radiology reports are a cornerstone of modern healthcare, capturing intricate diagnostic insights derived from medical images. These unstructured reports encapsulate the clinical context, imaging techniques, findings, and interpretations, which are pivotal in guiding patient care decisions.[1] However, their inherent lack of structure poses challenges for downstream applications that require standardized and structured data, including research, billing, accreditation, and quality improvement.[2] There is a push toward using structured formats instead of free-text radiology reports. Although initiatives such as RadReport[2] and

You may cite this article as: Abdullahi AA, Ganiz MC, Koç U, Gökhan MB, Aydın C, Özdemir AB. Deep learning for named entity recognition in Turkish radiology reports. *Diagn Interv Radiol.* 28 February 2025 DOI: 10.4274/dir.2025.243100 [Epub Ahead of Print].

RadLex[3] have helped standardize radiology reporting, unstructured formats remain the most common format despite the need for standardization. Various research methodologies have been investigated to bridge this gap, including rule-based systems, machine learning, and deep learning.

Our study focuses on applying deep learning to extract named entities from radiology reports written in Turkish. In addition, we developed a new dataset to train the named entity recognition (NER) task and considered the distinctive characteristics of the Turkish language to attain the best possible results. For the NER task, we utilized the DYGIE++ framework[4] and adapted it to the Turkish language. The DYGIE++ framework relies on a bidirectional encoder representations from transformers (BERT)[5] model to extract text embeddings. Therefore, we used the BioBERTurk model,[6] a variant of BERT pre-trained on Turkish biomedical data. This combination allows for the extraction of structured information, which can be used to enhance various medical applications. Our approach builds on previous research and aims to improve the overall effectiveness of information extraction in radiology reporting.

The potential of deep learning applications in Turkish radiology reports has yet to be fully explored. To remedy this, we worked with Ankara Bilkent City Hospital radiologists and hand-labeled a substantial dataset of 1,056 reports. To the best of our knowledge, this is the first dataset in Turkish created for this purpose. These reports have been expertly labeled to include observation and symptom categories, and they serve as a crucial foundation for our experiments.

In this paper, we provide a detailed explanation of our methodology and showcase how using DYGIE++ with various BERT models has been effective for our NER task of extracting observations and symptoms from Turkish radiology reports. Although there are no studies against which we can compare our F1 results (80.1) in Turkish, our results are similar to those in other languages. The implications of our study go beyond Turkish radiology reports; the lessons we learned and the methodologies we established can be applied to multiple languages and medical contexts, leading to improved information extraction practices. We hope to see a future where structured insights can be easily extracted from unstructured reports, leading to a revolution in medical reporting practices.

Main points

- Precise data are extracted from radiology reports to address the challenges of retrieving information from unstructured reports.
- Named entity recognition is used to identify observations and symptoms, even in low-resource languages such as Turkish.
- Diagnostic precision is improved and decision-making expedited to foster improved patient care and healthcare outcomes.

In the following sections, we will present related research and discuss the methodology, results, and conclusions that support our findings. The methodology section will elaborate on the dataset and experimental setup. In the results section, we will showcase the findings of our experiments conducted using varying configurations. In the discussion, we will compare our results with other studies in the field, including those conducted in languages other than Turkish. We hope to contribute to the ongoing dialogue on integrating deep learning into radiology reporting and inspire innovation in healthcare.

Methods

We created a labeled dataset of 1,056 radiology reports produced by the radiologists in our research team. Due to ethical and privacy considerations, it was not feasible to use actual patient data. Therefore, the radiologists drew from their experience of composing authentic radiology reports to generate synthetic reports that resembled the structure and content of genuine ones. This approach ensured that the dataset retained the critical features and complexities of real reports while safeguarding patient confidentiality and data privacy. The reports focused on computed tomography (CT) scans of the thorax area, encompassing the chest, lungs, heart, abdomen, and other vital organs. Figure 1 shows an example of a labeled report. This dataset can be utilized in various medical research projects and assist in developing diagnostic tools and techniques. Table 1 enumerates imaging types and their frequencies. We used the expertise of radiologists to label the data for NER, resulting in nine labels: Obs_Present, Obs_Uncertain, Obs_Technical, Obs_Anatomy, Obs_Absent, Obs_Advice, Symptom_P, Symptom_A, and Differential_Diagnosis. Table 2 enumerates the labels, their descriptions, and their frequencies.

We established a Doccano platform to simplify the labeling of our reports. Doccano is an open-source web-based annotation tool that provides a collaborative environment for annotating text elements such as named entities. It allows users to upload text documents and add annotations to a group of words within the document. Users work in parallel on separate documents that need to be labeled. Due to its user-friendly interface, Doccano was particularly valuable in simplifying the labeling process. An export of the data to the JavaScript object notation lines format became readily available once the labeling was complete. We labeled 1,056 reports by randomly dividing them into three equal parts for three radiologists to label in parallel. Ural Koç, co-author reviewed the labeling results and supervised the entire labeling process.

We named our task entity recognition using the DYGIE++ framework. The DYGIE++ framework is a span-based model for extracting entities, relations, and event triggers. We performed the entity extraction in isolation to accomplish our task. Our approach in the four stages of the DYGIE++ model is detailed as follows:

1. Token encoding: This step uses a BERT model to obtain token representations of the text. It utilizes a sliding window technique, feeding a sentence to the model at each iteration along with 15 surrounding sentences. We experimented with four BERT models: BERTurk, BioBERTurk, PubMedBERT, and XLM-RoBERTa. The BERTurk model was pretrained on Turkish text sourced from Wikipedia dumps, and we selected it because the model's Turkish language matched our training data. The BioBERTurk model was pretrained on top of BERTurk with Turkish biomedical texts and radiology theses, making it the most suitable fit for our application domain and language. The PubMedBERT model was pre-trained on English text sourced from the abstracts and articles of academic biomedical publications, and we selected it because its medical data matched our domain. The XLM-RoBERTa model was pre-trained on text from Wikipedia dumps containing 100 languages (including Turkish), and we chose it because medical terms tend to remain consistent across multiple languages.

2. Adaptive span enumeration: A span is a group of adjacent tokens that can be either a single token or a combination of many. We created it by concatenating token representations. The usage of suffixes in the Turkish language results in shorter sentences despite longer word lengths. For instance, the English phrase "the nasogastric tube has been pushed forward" translates to "nazogastrik tüp ileri itildi" or "nazogastrik tüp iler-

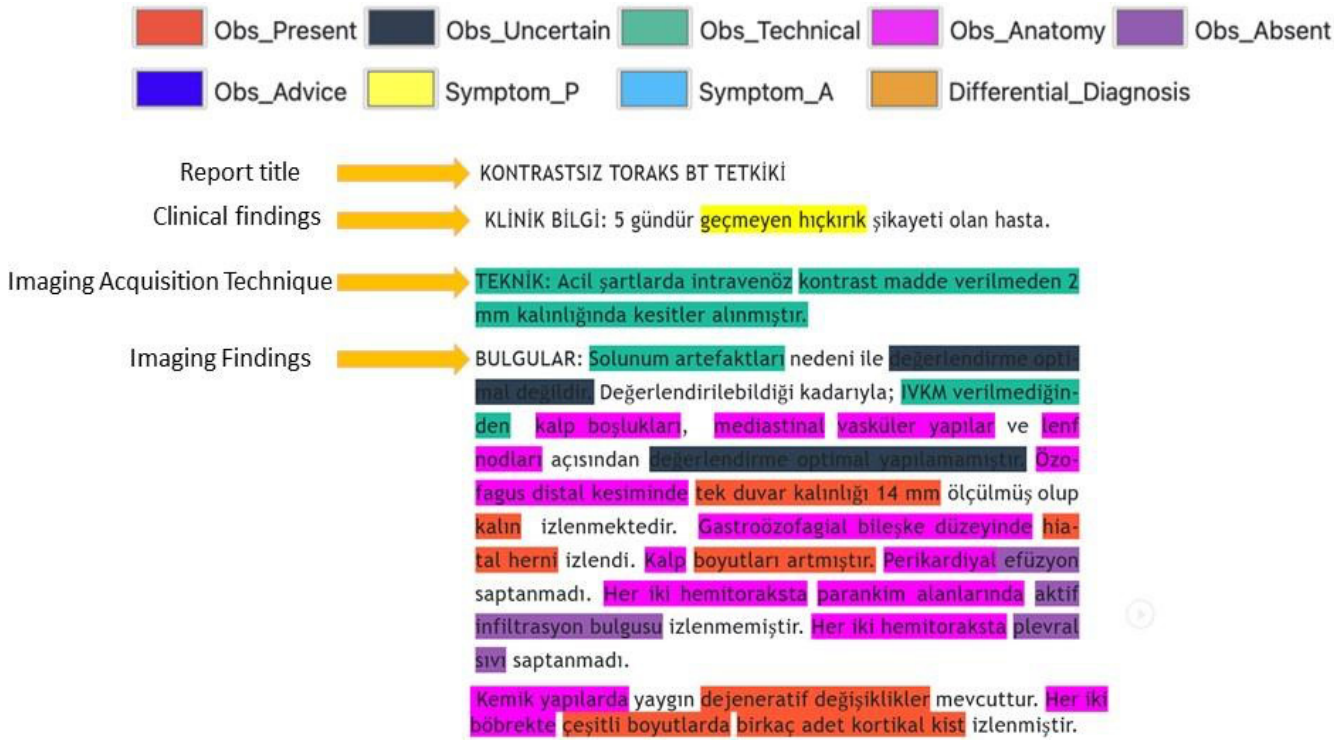

Figure 1. Color-coded example of labeled reports.

Table 1. Imaging types and their frequencies in the labeled dataset of radiology reports

Imaging type	Number of reports	Percentage
Abdominal radiology	363	34.38%
Thorax radiology	224	21.21%
Neuroradiology	187	17.71%
Vascular and thorax radiology	101	9.56%
Musculoskeletal radiology	66	6.25%
Head and neck radiology	45	4.26%
Vascular and thoracoabdominal radiology	25	2.37%
Vascular and neuroradiology	22	2.08%
Vascular and musculoskeletal radiology	15	1.42%
Vascular and abdominal radiology	5	0.47%
Vascular and neck radiology	3	0.28%

letildi" in Turkish, consisting of four- or three-word sentences instead of the seven-word sentence in English. Although DYGIE++ was originally developed using English, we modified our model to accommodate Turkish. We set the maximum number of tokens per span to four instead of the default limit of eight used in English experimentally, as we obtained the best performance using this value.

3. Span graph propagation: This step generates a multidirectional graph by computing the connections between spans. Spans are considered connected if they are likely to be related or refer to the same topic (coreference). We were interested in the coreference propagation in this step, which is crucial for identifying references to an entity throughout the document. Therefore, once we had the entity type of one reference, we could apply it to all the other references in the document.

4. Named entity classification: In the final step, a two-layered feed-forward neural network was used as a scoring function to make predictions for named entities.

For the experiment, we partitioned the 1,056 reports in the labeled dataset into three subsets: 75% for training, 15% for testing, and 10% for development. The training configuration closely followed that of DYGIE++.[4] The training phase spanned 100 epochs and focused on NER; therefore, the loss weights for relation extraction, coreference resolution, and event extraction were set to 0, and the weight for NER was set to 1. We used the AdamW optimizer,[7] with a learning rate of 1e −3 and weight decay of 0.0. The gradient norm was set to 5.0 for stable training with a slanted triangular learning rate scheduler. We used an NVIDIA V100 graphical processing unit as a CUDA device throughout the experiments. The codebase was in Python. We sourced our code from the DYGIE++ GitHub code repository of[4] (github.com/dwadden/dygiepp), which was built on the AllenNLP framework.[8] The loss weights are given as NER: 0.5, relation extraction: 0.5, coreference resolution: 0.0, and event extraction: 1.0.

Statistical analysis

As for the statistical analysis, we used the micro F1 score as the standard to evaluate and compare the performance of our models. Numeric values are given as a number and frequency (%). Cohen's kappa statistic was used to evaluate agreement. A *P* value <0.05 was considered statistically significant. The study did not require ethics committee approval or patient consent.

Results

Our setup comprises four experimental combinations differentiated by the BERT model, as described under "Token encoding" in section 2. Table 3 shows each model's precision, recall, and F1 score. The best perform-

Table 2. Label distribution for the dataset of radiology reports

Code	Name	Description	Frequency	Percentage
Obs_Present	Observations present	Presence of radiological features, identifiable pathophysiological processes, or diagnostic diseases	12,848	34%
Obs_Absent	Absence of observations	Absence of radiological features, identifiable pathophysiological processes, or diagnostic diseases	3,165	8.38%
Obs_Uncertain	Uncertain observations	Lack of certainty about a radiological feature, pathophysiological process, or diagnostic disease	1,102	2.92%
Obs_Technical	Technical observations	Technical situation that describes radiological techniques such as acquisitions	1,546	4.09%
Obs_Anatomy	Anatomical observations	Anatomical parts such as "vertebrae"	16,872	44.65%
Obs_Advice	Observations of advice	Tests and examinations recommended by the radiologist regarding the current diagnosis and treatment process	489	1.29%
Symptom_P	Presence of a symptom	A specific clinical symptom communicated by the clinician to the radiologist	668	1.77%
Symptom_A	Absence of a symptom	Absence of a specific clinical symptom communicated by the clinician to the radiologist	16	0.04%
Differential_ Diagnosis	Differential diagnosis	Differential diagnoses that may occur as a result of the current findings	1,084	2.87%

Table 3. Results for the BERTurk, BioBERTurk, PubMedBert, and XLM-RoBERTa models

BERT model	Precision	Recall	F1
BERTurk	78.3	79.9	79.1
BioBERTurk	**80.0**	**80.1**	**80.1**
PubMedBERT	75.0	76.9	75.9
XLM-RoBERTa	79.5	77.0	78.3

ing model was the BioBERTurk model, with an F1 score of 80.1. The BERTurk, PubMedBERT, and XLM-RoBERTa models scored 79.1, 75.9, and 78.3, respectively.

Figure 2 is a bar chart that displays each label's F1 score for all four BERT models. We report their respective precision, recall, and F1 scores using tables in Appendices 1-4. These tables offer a label-specific perspective, highlighting the strengths and weaknesses of each model. We can see that although the label "Obs_Present" is the most frequent (occurring 50.65% of the time), it does not have the highest F1 score among all the models. This affects the micro average F1 score because labels that occur more frequently contribute more weight to the overall F1 score. Conversely, "Symptom_A" has a 0.0 F1 score for all models because it lacks examples (only 16 occurrences) for the model to learn. Consequently, its effect on the overall F1 score is negligible.

After receiving constructive feedback from the peer reviewers, two radiologists who were not involved in the initial study evaluated the synthetically generated reports using a Likert scale. The Likert scale ranged from 1 to 5, where 1 indicated the least resemblance to real-world reports and 5 indicated the closest resemblance. The responses were analyzed using Cohen's kappa statistic (Cohen's kappa score: 0.92, $P < 0.001$). The evaluation of radiology reports prepared by the study radiologists achieved a high inter-observer agreement among the independent radiologists. Furthermore, the selected categories on the scale indicated that the reports closely resembled real-world radiology reports (Figure 3). After the peer-review process, 25% of the data were randomly re-annotated (UK) to assess intra-annotator agreement. Cohen's kappa statistic was used to evaluate the level of agreement, yielding a kappa value of 0.997 ($P < 0.0001$). This result indicates a high level of agreement and is statistically significant.

The co-occurrence chord diagram and matrix of the nine labels are shown in Figures 4, 5 and Appendices 5, 6.

Discussion

Structured reports have a standardized language and are consistently organized into ordered sections to enable the automated or semi-automated abstraction of reporting data. In recent years, numerous researchers have demonstrated a keen interest in extracting information from unstructured radiology reports, as almost all reports are written in this format. In 2010, Soysal et al.[9] proposed a natural language processing (NLP) system that converts radiology reports into Turkish. The initial medical information extraction system in Turkish, TRIES, follows a three-step conversion process. It begins with a morphological analysis of every word in the sentence, followed by NER and relation extraction. Its purpose is to match the sentence with a set of rule templates. An example is the sentence "The liver is 14 cm in height," which is analyzed as "Liver vertical tall + NESS + POSS3SG 14 cm + COP," later transformed into "[entity: Liver] [attribute: height] + POSS3SG [value: NUMERIC: 14 cm] + COP," and finally converted to "Liver.height = 14cm." The TRIES system has achieved results with a 93% recall and 98% precision rate. However, this method is limited because rule-based systems fail if a relationship cannot be matched to a specific rule.

Little research related to the present study has been conducted in the Turkish language domain. This is a notable shortcoming considering the considerable advancements published in the literature, especially in pre-trained deep learning models. One of the most commonly used pre-trained language models for creating downstream NLP applications via fine-tuning is BERT, which considers the entire context of words by look-

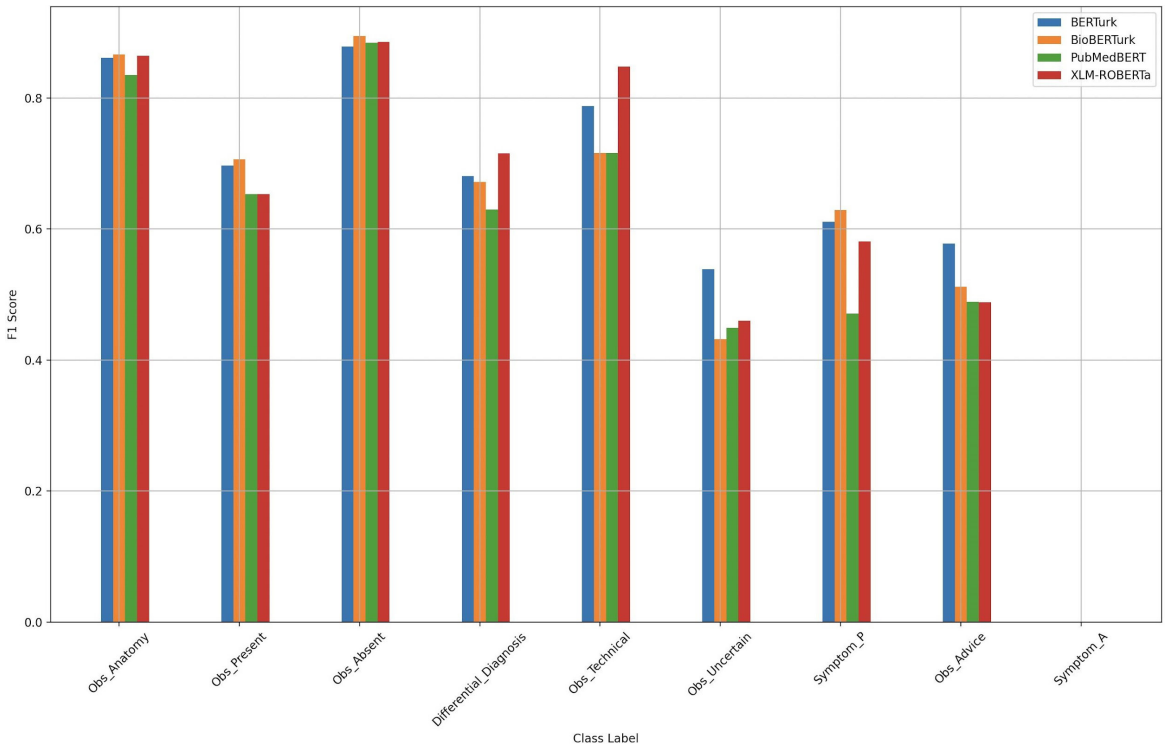

Figure 2. Bar chart of the F1 scores of the BERTurk, BioBERTurk, PubMedBert, and XLM-RoBERTa models.

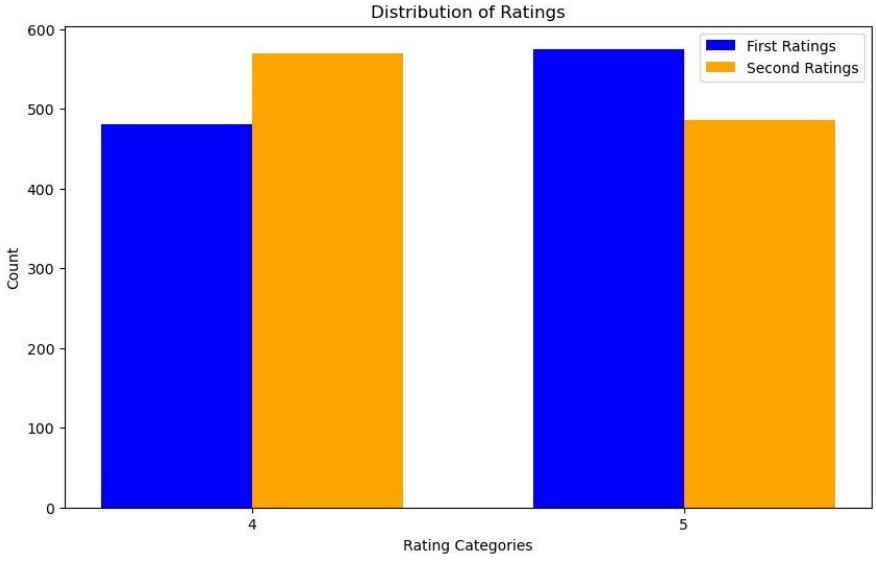

Figure 3. Evaluation of synthetically generated radiology reports by two independent radiologists using a Likert scale (1-5). The analysis showed a high inter-observer agreement (Cohen's kappa score: 0.92, $P < 0.001$), with the majority of scores indicating a strong resemblance to real-world radiology reports.

ing both left and right in a sentence. This model's innovation lies in its pre-training process, which is trained on large amounts of data to perform two tasks. First, masked language modeling (MLM) requires masked words within sentences to be predicted, helping BERT understand the word context and semantics. Second, next sentence prediction (NSP) predicts if one sentence follows another, enabling BERT to grasp sentence relationships. With this bidirectional approach, MLM and NSP allow BERT to capture intricate language relationships. In addition, BERT's architecture allows it to be fine-tuned for specific language-related tasks such as NER. As our task is NER on Turkish data, we experimented with four variations of BERT: BERTurk,[10] BioBERTurk, PubMedBERT,[11] and RoBERTa-XLM.[12]

We found no previous studies related to deep learning in the Turkish language; therefore, we explored other underrepresented languages to gain inspiration to help fill this gap. In a recent study, Jantscher et al.[13] investigated methods for NER and relation extraction from radiology reports in German. To achieve their goal, they fine-tuned a BERT model and used active learning for domain adaptation and training. Three separate datasets were utilized in this study. Reports on head CT were used to fine-tune the GermanMedBERT[14] model, and reports on magnetic resonance imaging (MRI) of the head and pediatric X-rays were used for domain adaptation and training. The researchers aimed to demonstrate that domain adaptation and active learning enhance the effectiveness of NER and relation extraction tasks. The model trained on MRI data performed the best, with an F1 score of 86.0 for NER and 80.0 for relation extraction.

In a similar study,[15] researchers aimed to extract named entities from Polish radiology reports. Using a dataset of 1,200 chest X-ray reports, the study focused on sequence labeling using the inside–outside–beginning annotation schema. This annotation schema consists of 44 tags representing everyday radiological observations while emphasizing generalization for potential application across clinical domains. The experiments involved the use of five BERT models: Pol-

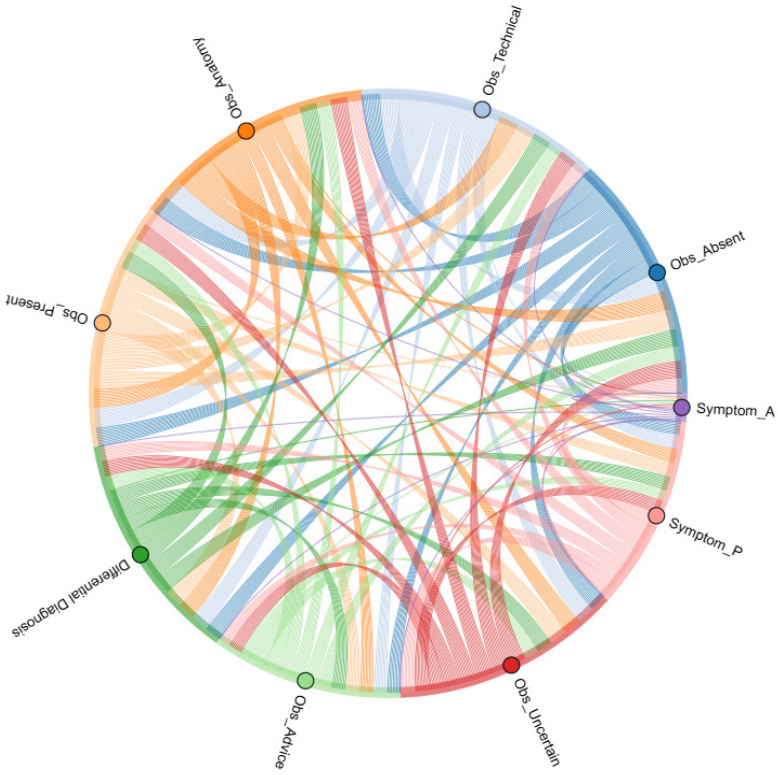

Figure 4. Co-occurrence chord diagram representing the total number of times each label pair appeared together across all reports. In this case, repetitions within the same document are also considered.

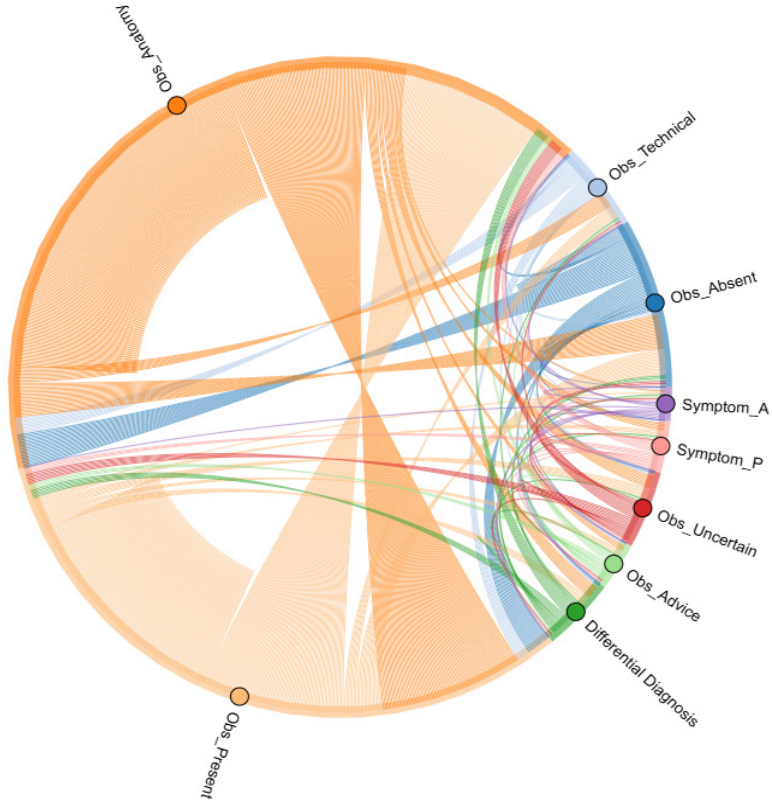

Figure 5. Co-occurrence chord diagram representing the frequency of unique label pairs. Even if the same label pairs appear multiple times, they are counted only once, illustrating the occurrence frequency of these unique combinations.

ish-roberta-base-v2,[16] Polish-distilroberta,[16] Polish-longformer,[16] HerBERT,[17] and mLUKE.[18] The mLUKE model is a multilingual variant of the LUKE model,[19] whereas the rest of the models were pre-trained only on Polish data. The results demonstrated that mLUKE was the most effective model, with an F1 score of 80.9. Its multilingual nature enhanced the domain-specific medical knowledge base across all languages. Certain classes exhibited lower-than-expected scores due to the complexity and variability within those categories. By contrast, others performed well despite limited annotated examples.

Another study[20] focuses on NER applied to chest CT reports in Japanese. The dataset consists of 118,155 reports, 540 of which were annotated by medical experts. Three deep learning models (BiLSTM-CRF, BERT, and BERT-CRF) were used to train NER. Each of the three models was pre-trained on Wikipedia data and CT reports. The labeled dataset was used to evaluate the models, which showed promising results in extracting clinical information from the Japanese chest CT reports. The BiLSTM-CRF model had the highest micro F1 score, with 95.4 for CT and 94.3 for Wikipedia. Higher F1 scores were observed across all models when pre-training with CT reports instead of only Wikipedia. Analysis of the effect of various modifiers on performance shows that the "certainty modifier" entity had a favorable impact, resulting in higher F1 scores. Conversely, the "change modifier" and "characteristics modifier" entities reduced performance, leading to lower F1 scores.

The results of the present study demonstrate that, among the different BERT models, BioBERTurk performed the best. We attribute our model's improved performance to adaptive span enumeration. We ran several iterations to determine the optimal value for the maximum number of tokens per span for the Turkish language. We set it at four instead of the default value of eight in English experimentally, as detailed in the Material and Methods section under "Adaptive span enumeration." This estimation resulted in a 1.5-point increase in BioBERTurk's F1 score. We believe this value to be specific to the Turkish language, and a similar concept can be applied to other languages.

The BERTurk model closely followed BioBERTurk in performance (79.1%) due to its Turkish language embeddings. This is a BERT model that was pre-trained from scratch using only Turkish text. Therefore, we expected it to perform better than multilingual models

such as XLM_RoBERTa and English-only models. The XLM-RoBERTa model performed reasonably well (78.3%), but, as expected, it was too generic because it was trained on data from 100 languages. In addition, it is a much larger model, and given its size, we needed a larger dataset for fine-tuning to have a noticeable impact. Finally, PubMedBERT is an English-only model that is pre-trained using English-only medical domain texts. Although the medical terminology in English and Turkish overlaps to a certain degree due to the heavy use of Latin in medicine, as mentioned before, Turkish is very different from English, especially in terms of the heavy use of suffixes that can modify medical concepts in Latin. For example, "appendix" in English can be translated as "Apendiks," "Apendiksin," or "Apendiksinin," with the suffix "-in" indicating possession or a relationship. Similarly, "intubation" in English can be expressed as "Entübasyon," "Entübasyonu," with the suffix "-u" for possession, or "Entübasyonunda," with the locative suffix "-da" to indicate location within a procedure, and so on. There can be a large number of variations with many different suffixes. Due to these profound differences between languages, we observed a significant drop in performance (80.1% vs. 75.9%) when we used PubMedBERT. For PubMedBERT, fine-tuning the model with a large number of Turkish medical texts may increase its performance. A possible solution for PubMedBERT to be considered in future studies is the use of adapters.[21] This method of fine-tuning adds extra layers to the model while retaining the existing ones, which are frozen during training. In this manner, the model preserves its medical knowledge by not updating the frozen weights and incorporates the Turkish context by updating the introduced weights. Our results indicate that it may be difficult to apply deep learning models that have been pre-trained on different languages or even multi-lingual models in domain specific applications such as medicine; however, it is worth using pre-trained models in the target language, adjusting hyper parameters, and applying domain specific fine-tuning.

Our resources, mainly medical data in Turkish, are limited due to the low number of datasets and studies. In fact, our dataset of 1,056 annotated radiology reports is a first in the Turkish medical domain. There are also restrictions for unlabeled data, both in terms of quantity and quality, in the Turkish medical domain compared with the English domain. These restrictions affect our model in several ways. First, we can discuss the domain specialization of large language models such as BERT. Although we used BioBERTurk as a base model that has been fine-tuned for the Turkish medical domain, we might obtain better results by further fine-tuning this model if we had access to a large number of anonymized Turkish radiology reports or related literature in Turkish. Second, we used just 1,056 Turkish radiology reports, which were manually created by radiology experts to mimic actual patient reports. This number can be increased in two ways. One is to involve more experts, which may not be feasible without vital funding and organization, currently beyond our capabilities. The other is to use techniques such as data augmentation,[22] which are useful for increasing the size of the labeled dataset, although the quality would be debatable. Furthermore, these medical text data augmentation methods are devised for English biomedical texts, and applying these directly to Turkish radiology reports may not be feasible due to the key differences between English and Turkish and the agglutinative nature of Turkish, as previously discussed.

We note that the four models exhibit different performance levels for each label. For instance, XLM-RoBERTa performs best for "Obs_Technical," as technical terms are not unique to the Turkish language and were pre-trained in multiple languages. Moreover, BioBERTurk has excellent results for "Symptom_P," as it was trained on the relevant Turkish biomedical data. The "Obs_Uncertain" label posed challenges for all four models because uncertainties usually involve negation-related terms such as "could not be measured" or "evaluation is not optimal." Consequently, most of these predictions tend to be misclassified as "Obs_Absent." The BERTurk model demonstrated the best performance for this specific class label because it is specially trained for the Turkish language. However, the unexpected underperformance of BioBERTurk in predicting the "Obs_Uncertain" label is noteworthy, given its pre-training on Turkish biomedical data. This performance discrepancy warrants a closer examination of pre-training data specificity.

The F1 score of 89.0 for Polish radiology reports in[13] closely aligns with our obtained score of 80.1. The dataset sizes are similar; ours has 1,056 instances, whereas theirs has 1,200. In addition, as in our study, certain classes are high frequency and yield lower F1 scores. We believe that the limited linguistic resources in both the Polish and Turkish languages specific to radiology reporting are the reason for this commonality. The F1 score reported by Sugimoto et al.[20] on Japanese data exceeds ours, and this difference can be attributed to the substantial amount of fine-tuning data they used, totaling over 100,000 reports. In our study, we faced constraints in conducting extensive fine-tuning due to the limited data available. The discrepancy in fine-tuning resources underscores the impact of data volume on model performance and highlights the importance of considering the scale of training data in achieving optimal results. Despite these differences, we see parallel trends in the outcomes of certainty labels.

This study has several limitations that warrant consideration. First, the dataset used in this study was synthetic, created by radiologists to mimic actual Turkish radiology reports. This limitation could affect the generalizability of the findings to real-world applications. Second, a larger dataset, including actual anonymized reports, could enhance the robustness and performance of the models, particularly in identifying less frequent entity labels such as "Symptom_A." Third, although the study focuses on Turkish radiology reports, the findings may not be directly applicable to other low-resource languages without language-specific adaptations. Similar adjustments would be necessary for other languages with unique linguistic features. Fourth, despite the strong performance of the BioBERTurk model, the study was limited to evaluating only four BERT-based models. Exploring additional model architectures or integrating ensemble approaches could potentially yield improved results. Finally, due to resource constraints, fine-tuning was performed with limited training data. Access to a larger corpus of Turkish biomedical texts or radiology reports could further optimize the performance of the deep learning models.

In our future research, we plan to use the mentioned insights to propose a pretrained BERT model for biomedical applications in Turkish. We also plan to develop a language-specific approach to determine optimal token span lengths during adaptive span enumeration. These initiatives will enhance the accuracy and efficiency of information extraction models, as demonstrated in our research. Based on recent developments in artificial intelligence (AI), mainly in large language models, we also plan to experiment with these models, such as GPT-4o and Llama 3, and with different sized models, and compare their performances.

In conclusion, our study highlights the critical role of language-specific adaptations and domain-relevant fine-tuning in enhancing NER for Turkish radiology reports. The introduction of BioBERTurk and the adaptive span enumeration mechanism proved instrumental in achieving the highest performance among the tested models. By experimentally determining an optimal span length tailored to the Turkish language, we demonstrated the necessity of customizing hyperparameters to accommodate linguistic features such as agglutination and complex suffix structures. Furthermore, this research is built on the first-ever NER dataset derived from Turkish radiology reports, a resource labeled by radiology experts. This dataset not only reflects the unique linguistic and domain-specific challenges of Turkish but also lays the groundwork for future advancements in low-resource medical NLP. Our work also underscores the challenges posed by limited annotated datasets and the importance of future efforts in expanding high-quality medical text resources. By leveraging advances in large language models and further fine-tuning with domain-specific data, we aim to push the boundaries of information extraction in low-resource languages. Ultimately, this research contributes to the development of AI tools that streamline clinical workflows, improve diagnostic precision, and enhance patient care. We hope our research contributes to continued innovation that enables healthcare practitioners to access standardized and structured data to improve patient care.

Acknowledgments

We would like to express our gratitude to the two independent radiologists (Yasin Celal Güneş, Mehmet Numan Çolakoğlu) who contributed to this study by evaluating the reports and providing their valuable assessments.

We have provided our data, code, and model files as supplementary links: https://drive.google.com/drive/folders/1VWWRhxd-KkHTvSWD1KreEEAKRcofOxqbo

GitHub: https://github.com/BIGDaTA-Lab-AI/dygiepp-multilingual-radiology

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Kundeti SR, Vijayananda J, Mujjiga S, Kalyan M. Clinical named entity recognition: challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)*. 2016;1937-1945. [CrossRef]
2. Kahn CE Jr, Langlotz CP, Burnside ES, et al. Toward best practices in radiology reporting. *Radiology*. 2009;252(3):852-856. [CrossRef]
3. Langlotz CP. RadLex: a new method for indexing online educational materials. *Radiographics*. 2006;26(6):1595-1597. [CrossRef]
4. Wadden D, Wennberg U, Luan Y, Hajishirzi H. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint. arXiv:1909.03546*, 2019. [CrossRef]
5. Devlin J, Chang MV, Lee K, Toutanova K. Bert: pretraining of deep bidirectional transformers for language understanding. *arXiv preprint. arXiv:1810.04805*, 2018. [CrossRef]
6. Türkmen H, Dikenelli O, Eraslan C, Callı MV, Özbek SS. BioBERTurk: Exploring Turkish Biomedical Language Model¨ Development Strategies in a Low-Resource Setting. *J Healthc Inform Res.* 2023;7(4):433-446. [CrossRef]
7. Loshchilov I, Hutter F. Decoupled weight decay regularization. *arXiv preprint. arXiv:1711.05101*, 2017. [CrossRef]
8. Gardner M, Grus J, Neumann M, et al. Allennlp: a deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*. Published 2018. Accessed December 23, 2024. [CrossRef]
9. Soysal E, Cicekli I, Baykal N. Design and evaluation of an ontology-based information extraction system for radiological reports. *Comput Biol Med.* 2010;40(11-12):900-911. [CrossRef]
10. Schweter S. BERTurk - BERT models for Turkish. *Software*. [CrossRef]
11. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc*. 2021;3(1):1-23. [CrossRef]
12. Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale. *arXiv preprint. arXiv:1911.02116*, 2019. [CrossRef]
13. Jantscher M, Gunzer F, Kern R, Hassler E, Tschauner S, Reishofer G. Information extraction from German radiological reports for general clinical text and language understanding. *Sci Rep*. 2023;13(1):2353. [CrossRef]
14. Manjil Shrestha. Development of a language model for the medical domain. *PhD Thesis*, Hochschule Rhein-Waal, 2021. [CrossRef]
15. Obuchowski A, Klaudel B, Jasik P. Information extraction from Polish radiology reports using Language models. *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*. 2023;113-122. [CrossRef]
16. Dadas S, Perelkiewicz M, Poswiata R. Pretraining Polish Transformer-Based Language Models at Scale. *Artificial Intelligence and Soft Computing*, 2020. Springer International Publishing. pages 301-314. ISBN: 9783-030-61534-5. [CrossRef]
17. Mroczkowski R, Rybak P, Wroblewska A, Gawlik I. HerBERT: efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, April 2021, Kiyv, Ukraine. Association for Computational Linguistics, pages 1–10. [CrossRef]
18. Ri R, Yamada I, Tsuruoka Y. mLUKE: the power of entity representations in multilingual pretrained language models. *arXiv preprint. arXiv:2110.08151*, 2021. [CrossRef]
19. Yamada I, Asai A, Shindo H, Takeda H, Matsumoto Y. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*, 2020. [CrossRef]
20. Sugimoto K, Takeda T, Oh JH, et al. Extracting clinical terms from radiology reports with deep learning. *J Biomed Inform*. 2021:116:103729. http://doi.org/10.1016/j.jbi.2021.103729. [CrossRef]
21. Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP. *Proc Int Conf Mach Learn*. 2019;2790-2799. [CrossRef]
22. Issifu AM, Ganiz MC. A simple data augmentation method to improve the performance of named entity recognition models in the medical domain. *Proc 6th Int Conf Comput Sci Eng (UBMK)*. 2021;763-768. [CrossRef]

Appendix 1. Labels results for BERTurk model

Labels	Precision	Recall	F1
Obs present	0.717	0.697	0.706
Obs absent	0.890	0.899	0.894
Obs uncertain	0.559	0.352	0.432
Obs technical	0.708	0.723	0.716
Obs anatomy	0.849	0.885	0.866
Obs advice	0.478	0.550	0.512
Symptom P	0.688	0.579	0.629
Symptom A	0.000	0.000	0.000
Differential diagnosis	0.580	0.797	0.671

Appendix 2. Labels results for BioBERTurk model

Labels	Precision	Recall	F1
Obs present	0.702	0.691	0.697
Obs absent	0.883	0.874	0.879
Obs uncertain	0.560	0.519	0.538
Obs technical	0.787	0.787	0.787
Obs anatomy	0.846	0.876	0.861
Obs advice	0.520	0.650	0.578
Symptom P	0.647	0.579	0.611
Symptom A	0.000	0.000	0.000
Differential diagnosis	0.585	0.814	0.681

Appendix 3. Labels results for PubMedBERT model

Labels	Precision	Recall	F1
Obs present	0.673	0.635	0.653
Obs absent	0.884	0.884	0.884
Obs uncertain	0.500	0.407	0.449
Obs technical	0.708	0.723	0.716
Obs anatomy	0.805	0.866	0.835
Obs advice	0.440	0.550	0.489
Symptom P	0.533	0.421	0.471
Symptom A	0.000	0.000	0.000
Differential diagnosis	0.536	0.763	0.629

Appendix 4. Labels results for XLM-RoBERTa model

Labels	Precision	Recall	F1
Obs present	0.695	0.616	0.653
Obs absent	0.892	0.879	0.886
Obs uncertain	0.606	0.370	0.459
Obs technical	0.867	0.828	0.848
Obs anatomy	0.853	0.876	0.864
Obs advice	0.476	0.500	0.488
Symptom P	0.750	0.474	0.581
Symptom A	0.000	0.000	0.000
Differential diagnosis	0.628	0.831	0.715

Appendix 5. A co-occurrence matrix showing the total number of times each label pair appeared together across all reports. Repetitions within the same document are included in the calculations

	Obs_ Absent	Obs_ Technical	Obs_ Anatomy	Obs_ Present	Differential diagnosis	Obs_ Advice	Obs_ Uncertain	Symptom_P	Symptom_A
Obs_Absent	32694	14895	167727	125985	10489	4715	10774	7010	164
Obs_Technical	14895	6950	81788	62152	5077	2413	5763	3179	83
Obs_Anatomy	167727	81788	918308	708388	56819	26141	60585	35245	851
Obs_Present	125985	62152	708388	543428	43728	20306	46456	26866	697
Differential diagnosis	10489	5077	56819	43728	3898	1785	3549	2375	65
Obs_Advice	4715	2413	26141	20306	1785	752	1896	972	20
Obs_Uncertain	10774	5763	60585	46456	3549	1896	4258	2332	48
Symptom_P	7010	3179	35245	26866	2375	972	2332	2190	47
Symptom_A	164	83	851	697	65	20	48	47	2

Appendix 6. A co-occurrence matrix showing the frequency of unique label pairs. Each pair is counted only once, regardless of how many times it appears within or across documents

	Obs_ Absent	Obs_ Technical	Obs_ Anatomy	Obs_ Present	Differential diagnosis	Obs_ Advice	Obs_ Uncertain	Symptom_P	Symptom_A
Obs_Absent	0	332	334	334	316	254	304	246	15
Obs_Technical	332	0	332	332	315	253	303	245	15
Obs_Anatomy	334	332	0	334	316	254	304	246	15
Obs_Present	334	332	334	0	316	254	304	246	15
Differential diagnosis	316	315	316	316	0	244	288	234	15
Obs_Advice	254	253	254	254	244	0	232	182	10
Obs_Uncertain	304	303	304	304	288	232	0	223	13
Symptom_P	246	245	246	246	234	182	223	0	14
Symptom_A	15	15	15	15	15	10	13	14	0