Jurnal Informatika dan Rekayasa Perangkat Lunak ISSN 2656-2855
Vol. 8, No. 1, March 2026, Page. 51-58 e-ISSN 2685-5518

Comparison of BioBERT and DistilBERT for Named Entity Recognition on Indonesian Radiology Clinical Data

Nadia Eka Aprilia[1*], **Danang Wahyu Utomo**[2]
[1,2] Informatics Engineering, Computer Science Faculty, Dian Nuswantoro University
[*]Email: 111202214053@mhs.dinus.ac.id

Absract

Named Entity Recognition (NER) in Indonesian language radiology reports faces significant challenges due to the limited availability of labeled data for model training. This constraint is a major obstacle to developing an accurate medical information extraction system. Pseudo-labeling emerges as a potential solution by leveraging abundant unlabeled data to expand the training dataset without the need for time-consuming manual annotation. This study aims to compare the performance of two transformer models, BioBERT and DistilBERT, fine-tuned on pseudo-labeled data for extracting medical entities from Indonesian radiology reports. The research methodology encompasses three main stages text preprocessing and normalization, text alignment using regular expressions with BIO labeling, and model fine-tuning with a pseudo-labeling strategy. Model performance was evaluated using Precision, Recall, and F1-score metrics on an adapted radiology dataset. The results indicate that pseudo-labeling was effective in enhancing the performance of both models. DistilBERT achieved a higher accuracy of 96,4, while BioBERT reached 92.78%. Nonetheless, DistilBERT demonstrated superior computational efficiency with faster training time. This study provides valuable insight for selecting an optimal model architecture for NER tasks on Indonesian medical text, considering the balance between accuracy and computational efficiency.

Keywords: *Named Entity Recognition, Pseudo-labeling, BioBERT, DistilBERT, Radiology report*

INTRODUCTION

Radiology reports are a rich source of clinical information containing unstructured data that is critical for medical decision-making. *Named Entity Recognition* (NER) plays an important role in extracting medical entities such as anatomical locations, findings, and procedures from these reports (Lee *et al.*, 2020). However, the development of NER models for Indonesian faces significant challenges due to the limited availability of adequately annotated datasets, a problem also observed in other non-English languages, such as Spanish. A study by (Pérez-Díez *et al.*, 2021) reported that an NER-based method using a *BiLSTM-CRF* architecture can achieve high performance (*Recall* of 97.18%) in the de-identification of Spanish radiology texts, even with a relatively small training dataset, highlighting the potential of similar approaches for low-resource languages.

Previous studies have demonstrated various approaches to processing radiology texts. (Tsuji *et al.*, 2021) developed a *RadLex-based* model for extracting compound terms; however, it still relies on a limited dictionary. (Djati Prinantyo and Salam, no date) successfully improved *BioBERT* performance by integrating *BiLSTM* and *CNN-Char*, demonstrating the potential of hybrid architectures. Meanwhile, (Abdullahi *et al.*, 2025) proved the effectiveness of adapting BERT-based models for morphologically complex languages such as Turkish, which shares similarities with Indonesian.

Weak supervision and *pseudo-labeling* techniques have been proven effective in addressing the limitations of annotated data. (Wang *et al.*, 2022) successfully leveraged *weak supervision* on 39,099 ophthalmology records, while (Kuligowska and Kowalczuk, 2021) highlighted the effectiveness of *pseudo-labeling* for biomedical text classification. A similar approach was also successfully applied by (Huang *et al.*, 2021), who developed a model using 400 fully annotated articles.

A recent study by (Tay *et al.*, 2024) developed a comprehensive NER pipeline for metastasis inference from radiology reports, achieving an *F1-score* of 0.93 and outperforming larger models such as *RadBERT*. This success highlights the importance of an integrated approach that combines NER with assertion detection and relation extraction. On the other hand, (Rohanian *et al.*, 2024) focused

on developing lightweight transformer models such as *ClinicalBERT*, which maintain strong performance with high computational efficiency, making them highly suitable for resource-constrained settings. Meanwhile, (Arzideh *et al.*, 2025), In their comparative study, also found that encoder-based models like *BERT* still outperform *Large Language Models* (LLMs) in complex clinical entity extraction tasks.

Negation is a crucial aspect in the interpretation of radiology reports. (Su, Babore and Kahn, 2025) demonstrated the superiority of transformer-based models such as *CAN-BERT* over *rule-based* approaches in negation detection, improving the *F1-score* from 0.492 to 0.777. This finding is further supported by (Abadeer, 2020), who showed that *DistilBERT* is capable of detecting Protected Health Information with performance comparable to specialized models.

A study by (Steinkamp *et al.*, 2021) developed an *LSTM-based* architecture for extracting contextualized recommendations from radiology reports, achieving a token-level *F1-score* of 89.2%. Meanwhile, (Paul *et al.*, 2022) evaluated various features for a *CRF-based de-identification* model on *NSCLC* radiology reports, identifying n-grams, *prefix-suffix, word embeddings*, and *word shape* as the most effective features. (Rao, no date) evaluated several *BERT* variants on microbiology NER tasks, with *BioBERT* achieving an *F1-score* of 75.35 after *hyperparameter tuning*, highlighting the importance of the generalizability of *pre-training* corpora for domain-specific applications.

A recent study by (Sato *et al.*, 2024) demonstrated the effectiveness of utilizing *free-text* radiology reports for multi-organ anomaly detection in abdominal CT scans without manual annotation, using a deep learning pipeline that achieved an AUC of 0.886 on an external cohort. This approach integrates multi-organ segmentation models with information extraction schemes from radiology reports, highlighting the potential of leveraging routine clinical data for AI model development (Liu *et al.*, 2020) also successfully developed an NLP pipeline for Chinese radiology reports, achieving an *F1-score* of 93.00% on NER tasks, demonstrating the effectiveness of similar approaches for non-English languages. Meanwhile, (Lima-López *et al.*, 2025) released the CARMEN-I dataset, which contains annotated medical records in Spanish and Catalan, emphasizing the importance of non-English language resources for the development of clinical NLP systems. This dataset includes annotations for six types of clinical entities and personal health information, providing a valuable foundation for developing similar models in the Indonesian language.

(Kumar, Malla and Sharma, 2025), In their *BioBERT-RxReadmit* study, They demonstrated how *BioBERT* can be utilized for clinical text analysis to predict hospital readmissions. Meanwhile, (Cabrera *et al.*, 2024) explored multilingual transfer learning for NER in mammography reports, achieving an *F1-score* of 0.9102 despite using models pre-trained in different languages.

Based on the literature review, this study aims to:

1) Apply *pseudo-labeling* and a *weak supervision* technique to build an Indonesian radiology NER dataset.
2) Compare the performance of transformer-based models (*BioBERT*) and their lightweight variant (*DistilBERT*).
3) Develop an integrated pipeline that combines NER with negation detection and *relation extraction*.
4) Evaluate the effectiveness of the approach on a limited Indonesian radiology dataset.

LITERATURE REVIEW

Research in the field of *Named Entity Recognition* (NER) on medical texts continues to develop, particularly in efforts to improve the ability of models to recognize important entities in clinical data. Various approaches have been proposed, ranging from *rule-based* methods to *machine learning* and *deep learning* methods. In recent years, deep learning-based approaches have become the most dominant because they can handle the high complexity of medical language.

One widely used approach in NER is transformer-based models. These models, such as *BERT* and its various derivatives, can better understand the context of words compared to previous methods. One frequently used variant is *BioBERT*, a model pre-trained on biomedical data. There is also a lighter model, *DistilBERT*, which is a distilled version of *BERT* that is

smaller in size but still maintains adequate performance.

Several studies have shown that transformer-based models, especially those that are domain-adapted, have superior performance in NER tasks. Kumar et al. (2025) showed that *BioBERT* provides better results than regular BERT in recognizing entities in clinical texts. This is also supported by other studies stating that domain-adapted models are more effective in handling the complexity of medical language.

However, models like *BioBERT* have limitations, namely the need for relatively high computational resources. This can be an obstacle when applied in resource-limited environments. Therefore, lighter models like DistilBERT are increasingly used because they offer a balance between performance and efficiency.

In addition, a major challenge in developing NER for Indonesian is the limited availability of labeled data. To address this, techniques such as *pseudo-labeling* can be used to leverage unlabeled data. Research by Momoki et al. (2022) showed that this approach is quite effective, although it has been applied more often to image data than text.

For non-English languages, some studies have also shown that combining model-based approaches and domain knowledge (such as specialized lexicons) can improve results. For example, studies by Liu et al. (2020) on Chinese, as well as research by Salazar Cabrera et al. (2024), showed that transformer models can adapt well to other languages through fine-tuning.

Although much research has been conducted, most still focus on English or languages with greater data availability. Meanwhile, research on NER in Indonesian medical texts remains limited, especially those that combine *pseudo-labeling* techniques.

Based on this, this study attempts to fill the gap by comparing two models, *BioBERT* and *DistilBERT*, on Indonesian radiology texts. The main difference from previous studies is the use of *pseudo-labeling* to help overcome data limitations, as well as evaluation not only in terms of *accuracy* and *F1-score* but also computational efficiency. Thus, this study is expected to provide a more comprehensive picture of the performance and efficiency of models under limited data conditions.

METHOD

A. Data Collection

This study utilizes a dataset consisting of 172 radiology reports obtained from hospitals in Central Java, covering the period from January 2022 to December 2023. Data were collected through the hospital's Radiology Information System (RIS), adhering to strict ethical procedures and patient anonymization processes. The dataset includes Indonesian-language MSCT abdomen examination reports, containing both structured and unstructured text information.

B. Research Scheme

The research stages as shown in Figure 1.

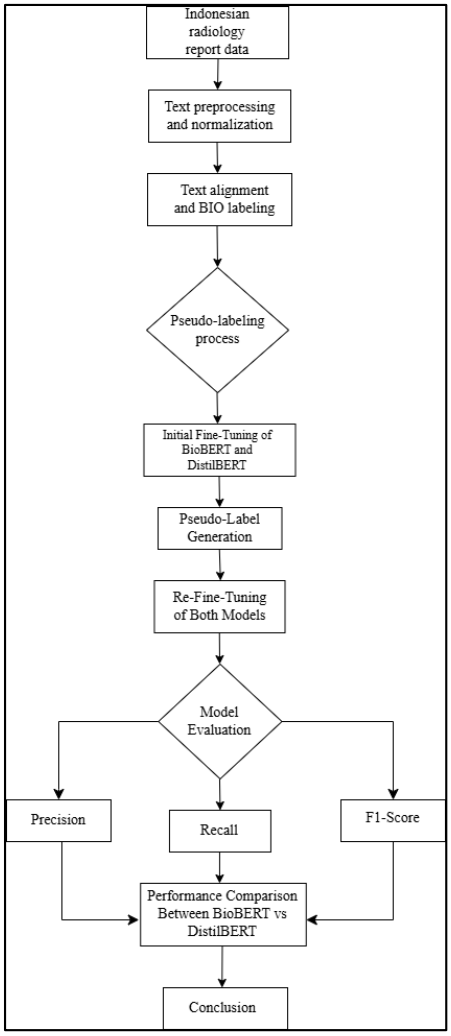

Figure 1 fine-tuning process flow for Indonesian clinical NER.

Based on Figure 1, the following is an explanation of the research stages according to

the flowchart, along with their sequence and process:

1) The data used consists of radiology texts written in Indonesian to train and test the models.
2) Before being processed by the model, the raw data must be cleaned and standardized. This includes *preprocessing* to remove punctuation, numbers, and special characters, as well as converting text to lowercase (*lowercasing*). Normalization is also performed to handle medical abbreviations, correct typographical errors, and standardize medical terms.
3) This stage involves identifying and classifying important entities in the text (e.g., anatomy, diseases, findings). Text alignment ensures that each word or sub word (token) in the text corresponds to the correct label. BIO labeling is a commonly used scheme for NER, where B-(Begin) indicates the beginning of an entity (e.g., B-DISEASE), I-(Inside) indicates the inside of an entity (e.g., I-DISEASE), and O-(Outside) indicates words that are not part of any entity.
4) A partially trained model is used to label (*unlabeled data*). The labels generated by this model are called *pseudo-label*.
5) Two (*pre-trained*) language models are (*fine-tuning*) using labeled radiology data. *BioBERT* is a *BERT-based* model pre-trained on biomedical text corpora in English and is capable of understanding medical terminology. *DistilBERT* is a more compact and faster version of *BERT*, trained on general-domain data.
6) After the initial *fine-tuning*, both models (*BioBERT* and *DistilBERT*) are used to predict labels on unlabeled data. These predictions are then used as *pseudo-label*, resulting in a larger augmented labeled dataset.
7) The dataset enhanced with *pseudo-label* is combined with the original labeled data. This larger combined dataset is then used for a second round of *fine-tuning* on both models (*BioBERT* and *DistilBERT*), with the aim of further improving model performance using more data.
8) After training is completed, the performance of both models is evaluated using a separate test dataset.
9) *Precision* measures, out of all entities predicted as positive by the model (e.g., "Disease"), how many are actually correct. High precision indicates that the model makes few false (*false positive*) errors.
10) *F1-score* is the harmonic mean of *Precision* and *Recall*. It serves as a single metric to evaluate the balance between the two, especially when the dataset is imbalanced.
11) *Recall* measures, out of all entities that should be predicted as positive, how many are successfully identified by the model. High *recall* indicates that the model misses few relevant entities (*false negative*).
12) At the evaluation stage, the results (*Precision, Recall, F1-score*) of both models are compared to determine which model performs better and is more suitable for implementation.
13) The final stage summarizes the entire process and findings.

C. Modeling

a) BioBERT

BioBERT is a transformer-based language model that has been specifically *pre-trained* on biomedical text corpora. The model was developed by further pre-training *BERT* base using large-scale biomedical datasets.

1) *BERT* base architecture with 12 encoder layers (110 million parameters).
2) Pre-training continued on biomedical corpora (PubMed abstracts + full articles).
3) Superior contextual understanding for medical terminology.
4) *Fine-Tuning* adds token-level classification layers for NER tasks.

b) DistilBERT

DistilBERT is a lighter and faster version of *BERT* developed using *knowledge distillation*. Its purpose is to retain much of *BERT* performance while significantly reducing size and increasing inference speed.

1) Distilled version architecture with 6 encoder layers (66 million parameters).
2) *Knowledge distillation* technique from *BERT* base.
3) Advantages 60% faster 40% smaller than *BERT* base.
4) The final stage, where the entire process and findings are summarized.

D. Evaluation

The performance evaluation of the *Named Entity Recognition* (NER) model is carried out by calculating standard metrics from the (*Confusion Matrix*) at the token level. The main metrics used are *Accuracy, Precision, Recall and F1-Score*.

1) True Positive (TP): Entities correctly predicted (span and label accurate).
2) True Negative (TN): Tokens that are not entities and are correctly predicted as non-entities.
3) False Positive (FP): Entities predicted but not present (over-prediction).
4) False Negative (FN): Entities present but not detected (under-prediction).

Formula and Explanation of Metrics:

1. Accuracy

The *accuracy* value is the ratio of the number of correct predictions to the total data. This metric provides an overall picture of how often the model makes correct predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FI} \quad (1)$$

2. Precision

Describes the *precision* level of the model from all data classified as positive; this metric answers the question "Of all predictions stating positive, what percentage are actually truly positive?" Precision is very important in cases where *false positives* must be minimized.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

3. Recall

Metrik This metric measures the model ability to retrieve all data that should be positive. *Recall* shows what percentage of the actual positives are identified by the model.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

4. F1-Score

This is the harmonic mean of *Precision* and *Recall*, providing a balanced perspective between both metrics. This metric is useful, especially when there is class imbalance in the dataset, as a high *F1-Score* can only be achieved if both Precision and *Recall* are high.

$$F1\text{-}Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

Process:

1) Evaluation is performed on a test set not used during *training*.
2) Calculations are done per entity type and macro-average.
3) *F1-Score* is the ultimate determinant of model performance.

RESULT AND DISCUSSION

A. Experiment Result

The *fine-tuning* experiments *BioBERT* and *DistilBERT* models for the *Named Entity Recognition* (NER) ask on Indonesian radiology report data were performed using a dataset that has been normalized and labeled according to the BIO scheme for five medical entity classes *Anatomy, Procedure, Measurement, Finding* and *Disease*. The models were evaluated using *Precision, Recall, F1-score, and Accuracy*.

a. Overall Model Performance

Table 1 is a summary of the performance of both models on the test data.

Tabel 1. Overall model performance

Model	Accuracy	F1-Score	Precision	Recall
BioBERT	92.78%	8.75%	8.43%	9.09%
DistilBERT	96.4%	21.4%	23.85%	20.9%

Based on the evaluation results shown in the classification table above, the *BioBERT* model achieved an *Accuracy* of 92.78%, with a *Precision* of 8.43%, *Recall* of 9.09%, and *F1-score* of 8.75%. Although the *accuracy* appears high, this is mainly due to the dominance of the non-entity label (“O”), which is much more prevalent than other medical entity labels. The model is only able to recognize the "O" token very well (*Precision* 0.9278, *Recall* 1.0000).

On the other hand, the *DistilBERT* model shows slightly better results in recognizing medical entities. This model achieves an *Accuracy* of 96.47%, with *Precision* 23.85%, *Recall* 20.98%, and *F1-score* 21.43%. Although still relatively low, *DistilBERT* managed to recognize some tokens with the finding label, especially in the B-finding class (*Precision* 0.9268, *Recall* 0.4130) and I-finding (*Precision* 0.7193, *Recall* 0.8952). This shows that *DistilBERT* has slightly better generalization capability toward linguistic patterns in Indonesian texts compared to *BioBERT*, even without explicit medical domain training.

The difference in performance indicates that domain-specific *pre-training* (as in *BioBERT*) does not always guarantee better results when the model is applied to a language different from its source language. *BioBERT* is trained on English medical corpora (PubMed and PMC), while *DistilBERT*, although not a medical domain model, is more flexible in recognizing general linguistic patterns also present in Indonesian.

Both models show the same challenge regarding (*class imbalance)*. The "O" label dominates more than 90% of tokens, causing the models to tend to ignore minority entities such as *disease* or *procedure*. This is reflected in the very low *F1-scores* for those labels.

B. Model Training and Validation Analysis

1) *BioBERT Loss Curve*

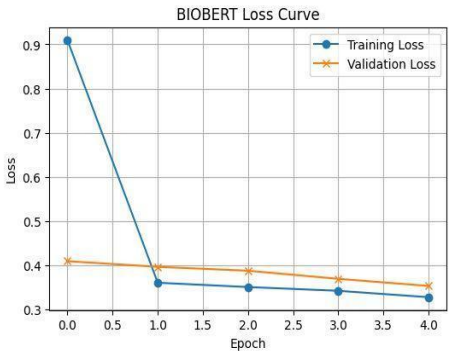

Figure 2 Fine-tuning *BioBERT*

In Figure 2, it can be seen that the *training loss* at the beginning of training (epoch 0) is quite high, around 0.9, and then drops sharply to about 0.33 at epoch 4. Meanwhile, the *validation loss* remains relatively stable in the range of 0.4–0.35 and decreases gradually as the epoch increase.

This condition shows that the *BioBERT* model experiences a stable training process without significant indications of *overfitting*, as the difference between the *training loss* and the *validation loss* is not too large. Although the loss value was successfully reduced, the classification evaluation results show that the model is not able to recognize medical entities well (low macro *F1-score*). This indicates that *BioBERT* merely learns to mimic the data distribution (especially the "O" label), but does not learn to represent the context of entities correctly due to domain limitations and language differences.

2) *DistilBERT Loss Curve*

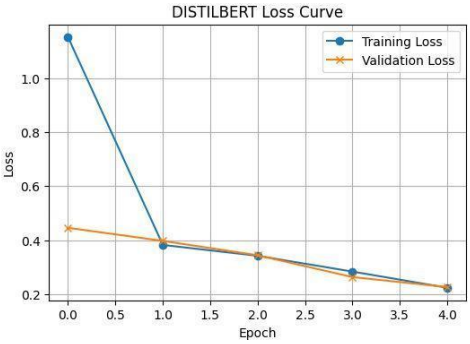

Figure 3Fine-tuning *DistilBERT*

In Figure 3, a similar downward pattern can be observed, but with a slightly higher initial *training loss* value (around 1.1) and a faster decrease, reaching 0.22 at epoch 4. Meanwhile, the *validation loss* also consistently decreases from 0.44 to 0.21, even slightly lower than the *training loss* at the end of *training*.

This indicates that *DistilBERT* has a more efficient and stable *training* process, with better generalization to the validation data. The fact that the *validation loss* approaches the *training loss* at the end of training shows that the model does not experience *overfitting* and is able to capture important patterns in the data.

Its higher classification performance (macro F1 0.2143 compared to *BioBERT* 0.0875) also strengthens the interpretation that *DistilBERT*, although not a medical domain model, is more adaptive to the linguistic

structure of the Indonesian language. The smaller model size and simplified architecture help *DistilBERT* adjust its representations with a limited dataset without losing stability during training.

CONCLUSION

Based on the experiments conducted, the DistilBERT model showed better performance compared to *BioBERT* on the *Named Entity* Recognition (NER) task for Indonesian clinical radiology texts. This is demonstrated by the F1-score of *DistilBERT*, which reached 0.2143, higher than *BioBERT* 0.0875. Although both models achieved high *accuracy* values (*DistilBERT* at 96.4% and *BioBERT* at 92.78%), these results were influenced by the dominance of non-entity tokens ("O"), making the *F1-score* a more representative indicator for evaluating model performance.

These results indicate that lighter models like *DistilBERT* are not only more computationally efficient, but also able to deliver better performance than models with domain-specific *pre-training* such as *BioBERT*, especially when applied to Indonesian-language data. Additionally, both models still face challenges in recognizing minority entities such as *diseases* and *procedures* due to data *distribution imbalance*.

For future research, the development of the annotation process can become a primary focus to produce higher-quality data. The involvement of expert annotators, such as radiologists, is expected to help build a more accurate gold standard dataset. Furthermore, the use of data augmentation techniques, *class imbalance* handling methods, as well as the exploration of multilingual models or models specifically adapted for the Indonesian language, also have the potential to improve model performance, especially in terms of *F1-score* and generalization capability.

REFERENCES

Abadeer, M. (2020) *Assessment of DistilBERT performance on Named Entity Recognition task for the detection of Protected Health Information and medical concepts*. Available at: https://github.com/huggingface/.

Abdullahi, A.A. *et al.* (2025) "Deep learning for named entity recognition in Turkish radiology reports," *Diagnostic and Interventional Radiology*, 31(5), pp. 430–439. Available at: https://doi.org/10.4274/dir.2025.243100.

Arzideh, K. *et al.* (2025) "From BERT to generative AI - Comparing encoder-only vs. large language models in a cohort of lung cancer patients for named entity recognition in unstructured medical reports," *Computers in Biology and Medicine*, 195. Available at: https://doi.org/10.1016/j.compbiomed.2025.110665.

Cabrera, E.R.S. *et al.* (2024) "Named Entity Recognition in Mammography Radiology Reports using a Multilingual Transfer Learning Approach," *Proceedings - IEEE Symposium on Computer-Based Medical Systems*. Institute of Electrical and Electronics Engineers Inc., pp. 273–277. Available at: https://doi.org/10.1109/CBMS61543.2024.00052.

Djati Prinantyo, G. and Salam, A. (no date) "Optimization of Biobert Model for Medical Entity Recognition Through Bilstm and CNN-Char Integration Optimalisasi Model Biobert untuk Pengenalan Entitas Medis melalui Integrasi Bilstm dan CNN-Char," 10(2), p. 2025.

Huang, D.-L. *et al.* (2021) "Accurate Name Entity Recognition for Biomedical Literatures: A Combined High-quality Manual Annotation and Deep-learning Natural Language Processing Study." Available at: https://doi.org/10.1101/2021.09.15.460567.

Kuligowska, K. and Kowalczuk, B. (2021) "Pseudo-labeling with transformers for improving Question Answering systems," *Procedia Computer Science*. Elsevier B.V., pp. 1162–1169. Available at: https://doi.org/10.1016/j.procs.2021.08.119.

Kumar, A., Malla, J. and Sharma, A. (2025) "Predictions Through Clinical Text Analysis with BioBERT," *International*

Journal on Engineering Artificial Intelligence Management, 02(2), pp. 14–29. Available at: http://creativecommons.org/licenses/by/4.0/.

Lee, J. *et al.* (2020) "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, 36(4), pp. 1234–1240. Available at: https://doi.org/10.1093/bioinformatics/btz682.

Lima-López, S. *et al.* (2025) "A textual dataset of de-identified health records in Spanish and Catalan for medical entity recognition and anonymization," *Scientific Data*, 12(1). Available at: https://doi.org/10.1038/s41597-025-05320-1.

Liu, H. *et al.* (2020) "A Natural Language Processing Pipeline of Chinese Free-Text Radiology Reports for Liver Cancer Diagnosis," *IEEE Access*, 8, pp. 159110–159119. Available at: https://doi.org/10.1109/ACCESS.2020.3020138.

Paul, T. *et al.* (2022) "Utility of Features in a Natural-Language-Processing-Based Clinical De-Identification Model Using Radiology Reports for Advanced NSCLC Patients," *Applied Sciences (Switzerland)*, 12(19). Available at: https://doi.org/10.3390/app12199976.

Pérez-Díez, I. *et al.* (2021) "De-identifying Spanish medical texts - named entity recognition applied to radiology reports," *Journal of Biomedical Semantics*, 12(1). Available at: https://doi.org/10.1186/s13326-021-00236-2.

Rao, B.K. (no date) *MICROBIAL NAMED ENTITY RECOGNITION USING BERT MODELS*.

Rohanian, O. *et al.* (2024) "Lightweight transformers for clinical natural language processing," *Natural Language Engineering*, 30(5), pp. 887–914. Available at: https://doi.org/10.1017/S1351324923000542.

Sato, J. *et al.* (2024) *Annotation-free multi-organ anomaly detection in abdominal CT using free-text radiology reports: a multi-center retrospective study*. Available at: www.thelancet.com.

Steinkamp, J. *et al.* (2021) "Automatic Fully-Contextualized Recommendation Extraction from Radiology Reports," *Journal of Digital Imaging*, 34(2), pp. 374–384. Available at: https://doi.org/10.1007/s10278-021-00423-8.

Su, Y., Babore, Y.B. and Kahn, C.E. (2025) "A Large Language Model to Detect Negated Expressions in Radiology Reports," *Journal of Imaging Informatics in Medicine*, 38(3), pp. 1297–1303. Available at: https://doi.org/10.1007/s10278-024-01274-9.

Tay, S.B. *et al.* (2024) "Use of Natural Language Processing to Infer Sites of Metastatic Disease from Radiology Reports at Scale," *JCO Clinical Cancer Informatics* [Preprint], (8). Available at: https://doi.org/10.1200/cci.23.00122.

Tsuji, S. *et al.* (2021) "Developing a RadLex-based named entity recognition tool for mining textual radiology reports: development and performance evaluation study," *Journal of Medical Internet Research*, 23(10). Available at: https://doi.org/10.2196/25378.

Wang, S.Y. *et al.* (2022) "Leveraging weak supervision to perform named entity recognition in electronic health records progress notes to identify the ophthalmology exam," *International Journal of Medical Informatics*, 167. Available at: https://doi.org/10.1016/j.ijmedinf.2022.104864.