SCIENCE PHOTONS Publishers

Research Article

Innovative Insights in Digital Health

Automated Natural Language Processing for Tumor Response Classification in Oncology Radiology Reports

Som Biswas*

Physician in Radiodiagnosis, Jefferson Abington Hospital, Pennsylvania, USA.

Received Date: 22 Jan 2026;
Accepted Date: 03 Feb 2026;
Published Date: 05 Feb 2026.

***Correspondence:** Som Biswas, MD, Physician in Radiodiagnosis, Jefferson Abington Hospital, Pennsylvania, USA.*

Citation: Som Biswas. Automated Natural Language Processing for Tumor Response Classification in Oncology Radiology Reports. Innov Insights Digit Health. 2026; 2(1): 1-3.

ABSTRACT

Natural language processing (NLP) has rapidly evolved in recent years, enabling the extraction of clinically relevant information from unstructured electronic medical records. Radiology reports, particularly in oncology, contain detailed longitudinal information on patient disease status, which can inform therapeutic decisions and outcomes. While structured reporting has been advocated to streamline data extraction, most radiology reports remain free-text [1]. This study aimed to leverage structured oncology reports (SOR) to train a deep NLP model for tumor response category (TRC) classification in free-text oncology reports (FTOR) and compare its performance with conventional NLP algorithms and human readers [2].

In this retrospective study, 9,653 SOR and 802 FTOR were analyzed from multiple radiology centers. A BERT-based NLP model was trained on SOR and applied to FTOR. Model performance was compared with radiologists, medical students, and radiology technologist students. The BERT model achieved an F1 score of 0.70, outperforming traditional NLP approaches and technologist students, approximating medical student performance, but was inferior to radiologists (F1, 0.79). Lexical complexity and semantic ambiguity reduced performance for both humans and machines [3].

***Conclusion:** Deep NLP models trained on structured oncology data can achieve near-human performance in extracting oncologic outcomes from free-text reports, offering a scalable approach for large-scale oncology data curation [4].*

***Keywords:** Natural Language Processing, Radiology Reports, Oncology Imaging, Tumor Response Classification, BERT, Structured Oncology Reports*

Introduction

Radiology reports in oncology provide a rich source of longitudinal data regarding tumor burden and treatment response. However, most reports are unstructured, posing challenges for systematic data extraction. Natural language processing (NLP) methods, ranging from rule-based algorithms to deep learning models, have emerged as powerful tools to retrieve information from free-text reports. Transformer-based architectures, such as BERT, have demonstrated superior performance in free-text classification tasks, enabling automated extraction of key clinical endpoints like tumor response, disease progression, and therapeutic outcomes [5].

Structured oncology reports (SOR) created during routine care can provide reliable ground truth labels for training NLP models. Using SOR to inform model training may bypass the need for extensive manual annotation, facilitating scalable data curation for multi-institutional studies.

Materials and Methods

Data Collection

Radiology databases from three independent centers were queried for CT, MRI, and ultrasound reports performed between March 2018 and August 2021. SOR were obtained from a tertiary care center, and FTOR were collected from a cancer research center and a hospital specializing in chest diseases. Reports lacking tumor assessment or duplicates were excluded [6].

Ground Truth Extraction

Structured oncology reports were mined for tumor response categories (progressive disease, stable disease, partial response, complete response) based on RECIST 1.1 criteria. Automated extraction employed rule-based NLP pipelines using regular expressions. Free-text oncology reports were manually annotated by experienced radiologists to provide ground truth labels.

NLP Model Development

Two types of NLP models were developed:

1. **Deep learning model:** BERT, fine-tuned on SOR oncologic findings.
2. **Conventional models:** Linear support vector classifier, k-nearest neighbors, and multinomial naive Bayes, using TF-IDF features.

The SOR dataset was split into training (85%) and test (15%) subsets. Fivefold cross-validation assessed model generalizability. Model performance was evaluated on FTOR using recall, precision, accuracy, and F1 score [7].

Human Comparison

Seven human annotators with varying expertise (radiologists, medical students, and radiology technologist students) independently classified TRCs from FTOR. Confidence levels were recorded on a five-point Likert scale.

Statistical Analysis

Performance metrics were calculated with 95% confidence intervals using bootstrap resampling. Intraclass correlation coefficients (ICCs) measured inter-rater agreement. Significance was set at $p < 0.05$.

Results

Patient and Report Characteristics

The final cohort included 10,455 patients (mean age 60 ± 14 years; 51% female). SOR were successfully mined for TRCs in 9,653 reports. FTOR datasets included 802 reports after exclusion criteria. Lexical complexity analysis showed FTOR varied in word count, vocabulary richness, and bigram usage [8].

Human Performance

Radiologists achieved an F1 score of 0.79, medical students 0.73, and technologist students 0.65. Inter-rater agreement was good to excellent among radiologists, moderate for medical students, and lower for technologist students. Confidence correlated positively with classification accuracy.

NLP Model Performance

The BERT model outperformed conventional NLP models, achieving an F1 score of 0.70 across all FTOR, with the highest AUC of 0.91 for concise disease-specific reports. Linear-SVC reached an F1 of 0.63, and other feature-rich models performed lower. Performance of NLP models decreased with increased lexical complexity and semantic ambiguity but tracked human performance trends.

Operational Use Case

TRC predictions by the BERT model enabled visualization of longitudinal tumor burden changes, demonstrating the potential for automated integration into tumor board assessments.

Discussion

Our findings demonstrate that NLP models trained on structured oncology data can accurately classify tumor response in free-text radiology reports. Deep learning models like BERT approximate human-level performance, outperforming conventional NLP approaches. Challenges remain in handling lexical variability and semantic ambiguity, highlighting the ongoing need for model interpretability and refinement.

By leveraging routinely generated SOR, large-scale automated curation of oncology outcomes is feasible, which may improve clinical decision support and facilitate multi-institutional research [9].

Conclusion

Deep NLP models can efficiently and accurately classify tumor response categories from free-text oncology reports. While performance does not surpass experienced radiologists, it reaches the level of medical trainees and offers a scalable approach to oncology data curation.

Conflicts of Interest

The authors declare no conflict of interest and received no specific funding for this work.

References

1. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008: 128–144.
2. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology. 2016; 279: 329–343.
3. Fink MA, Kades K, Bischoff A, Moll M, Schnell M, et al. Deep Learning-based Assessment of Oncologic Outcomes from Natural Language Processing of Structured Radiology Reports. Radiol Artif Intell. 2022; 4: e220055.
4. Kehl KL, Elmarakeby H, Nishino M, Van Allen EM, Lepisto EM, et al. Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. JAMA Oncol. 2019; 5: 1421-1429.
5. Johnson AWE, Pollard TJ, Shen L, Lehman LWH, Feng M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016; 3: 160035.
6. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.

7. Lee J, Yoon W, Kim S, Kim D, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020; 36: 1234–1240.

8. Liu Y, Ott M, Goyal N, Du J, Joshi M, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. 2019.

9. Casey A, Davidson E, Poon M, Dong H, Duma D, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak. 2021; 21: 179.