Research in Diagnostic and Interventional Imaging 6 (2023) 100027

Contents lists available at ScienceDirect

Research in Diagnostic and Interventional Imaging

journal homepage: https://www.journals.elsevier.com/redii

Original article

BERT-based natural language processing analysis of French CT reports: Application to the measurement of the positivity rate for pulmonary embolism

Émilien Jupin-Delevaux[a,*], Aissam Djahnine[b,c], François Talbot[d], Antoine Richard[d], Sylvain Gouttard[a], Adeline Mansuy[a], Philippe Douek[a,b], Salim Si-Mohamed[a,b], Loïc Boussel[a,b]

[a] Radiology department, Hospices Civils de Lyon - HCL, Lyon, France
[b] CREATIS, Univ Lyon, INSA–Lyon, Université Claude Bernard Lyon 1, UJM-Saint Etienne, CNRS, Inserm, CREATIS UMR 5220, U1294, Lyon, France
[c] Philips Research France, Suresnes, France
[d] DSN, Hospices Civils de Lyon - HCL, Lyon, France

ARTICLE INFO

Article History:
Received 23 August 2022
Accepted 15 March 2023
Available online 27 March 2023

Keywords:
Natural language processing
CT
Pulmonary embolism

ABSTRACT

Rationale and objectives: To develop a Natural Language Processing (NLP) method based on Bidirectional Encoder Representations from Transformers (BERT) adapted to French CT reports and to evaluate its performance to calculate the diagnostic yield of CT in patients with clinical suspicion of pulmonary embolism (PE).
Materials and methods: All the CT reports performed in our institution in 2019 (99,510 reports, training and validation dataset) and 2018 (94,559 reports, testing dataset) were included after anonymization. Two BERT-based NLP sentence classifiers were trained on 27.700, manually labeled, sentences from the training dataset. The first one aimed to classify the reports' sentences into three classes ("Non chest", "Healthy chest", and "Pathological chest" related sentences), the second one to classify the last class into eleven sub classes pathologies including "pulmonary embolism". F1-score was reported on the validation dataset. These NLP classifiers were then applied to requested CT reports for pulmonary embolism from the testing dataset. Sensitivity, specificity, and accuracy for detection of the presence of a pulmonary embolism were reported in comparison to human analysis of the reports.
Results: The F1-score for the 3-Classes and 11-SubClasses classifiers was 0.984 and 0.985, respectively. 4,042 examinations from the testing dataset were requested for pulmonary embolism of which 641 (15.8%) were positively evaluated by radiologists. The sensitivity, specificity, and accuracy of the NLP network for identifying pulmonary embolism in these reports were 98.2%, 99.3% and 99.1%, respectively.
Conclusion: BERT-based NLP sentences classifier enables the analysis of large databases of radiological reports to accurately determine the diagnostic yield of CT screening.

1. Introduction

In the era of personalized medicine, appropriateness and the adequacy of radiological requests is a major concern, particularly for ionizing examinations such as CT. In this context, professionals regularly publish recommendations for the proper use of CT examinations [1–3]. Likewise, many academic teams and private companies are focusing on developing automated decision support systems in order to reduce unnecessary imaging requests [4,5]. Measuring the effectiveness of these methods and recommendations is often difficult in practice. Indeed, it requires not only evaluating variations in the number of examinations requested but also their positive yield rate. The latter is complicated to acquire insofar as it requires an in-depth analysis of the examination reports. The use of structured reports or their indexing by pathology would make it possible [6], but these methods are seldom used in clinical practice, particularly in emergency settings.

A wide range of Natural Language Processing (NLP) methods have been used for years to automatically analyze unstructured CT reports and detect reported pathological findings [7–12]. This includes rule-based, machine learning-based, and hybrid algorithms. More recently, Bidirectional Encoder Representations from Transformers (BERT) [13], a new language representation model, has been introduced, which has state of the art performances for biomedical text mining [14,15]. Furthermore, BERT allows unsupervised domain

* Corresponding author.
E-mail address: emilien.jupin-delevaux@chu-lyon.fr (É. Jupin-Delevaux).

https://doi.org/10.1016/j.redii.2023.100027

adaptation through language model fine-tuning which makes it both particularly suitable for specific language domains such as clinical or radiological and adaptable to foreign languages [16,17].

In this study, we sought to develop and validate a BERT-based NLP method to extract the most common pathological features from chest CT reports and to evaluate its performance for pulmonary embolism (PE) detection in order to calculate diagnostic yield of CT in large academic hospitals in France.

2. Materials and methods

2.1. Population and database

All reports of CT examinations carried out at our institution between January 1, 2018 and December 31, 2019, were acquired from our radiological information system. CT reports from 2019 (99,510 reports, defined as training dataset) were used for the training and validation of the NLP model. Reports from 2018 (94,559 reports, defined as test dataset) were used for the PE detection and positive yield rate calculation part of our study. These reports were written by a very large and diverse panel of general and specialist radiologists in the Hospices Civils de Lyon, second teaching hospital in France with 13 public and they were structured according to the recommendations of the European Society of Radiology [6] in four parts: “Clinical referral”, “Technique”, “Findings” and “Conclusion”. Data usage policy of the Hospices Civils de Lyon in terms of confidentiality, anonymization and security was applied for each report and an IRB approval was obtained from the national ethical committee. Patients’ informed consent was waived by the ethical committee.

2.2. Training and validation of the classifiers

The training/validation process included two stages: 3-Classes and 11-SubClasses classification (described below). To train our two multiclass models, we first extracted all the sentences from the “Findings” section of each CT report from the 2019 training dataset. A random shuffle and a split were performed to create the training and the validating dataset with respectively 95% and 5% of the dataset. It resulted in 1,007,000 independent sentences where their entire manual labeling is virtually impossible. Among these sentences, 27,731 were randomly selected and labeled by two radiologists (EJD, LB) according to their semantic meaning. In case of disagreement between the two radiologists on a label, a review was performed in consensus.

Instead of performing pathology-specific whole-sentence classification, the first step of our proposed pipeline performs 3-Classes sentence classification, distinguishing only between "Non chest” (11277), "Healthy chest“(3808) and "Pathological chest" (12646) related sentences rather than particular pathologies. The second step used the “Pathological chest” labeled sentences to perform 11-SubClasses sentence classification indicating pathology type in a chest CT : “Pulmonary embolism” (560), “Non-embolic pulmonary artery pathology” (656), “Aortic pathology” (545), “Heart abnormality” (1215), “Mediastinal lymphadenopathy” (1161), “Lung consolidation” (1686), “Lung nodule or mass” (2201), “Bronchial disorders” (621), “Interstitial syndrome” (1322), “Pleural abnormality” (1216) and “Chest wall abnormality” (1463). A Sankey diagram of these labels is provided in Figure 1.

A NLP model based on Transformers was trained and evaluated in two stages. The first stage consisted in an unsupervised fine-tuning of the CamemBERT French pretrained model (‘camembert-large’) of BERT using a Masked Language Modeling [18]. Typically, to achieve this task, tokenization is the very first step in most NLP tasks, the process of grouping and segmenting text into meaningful chunks in words or sub-words level. A token serves as an atomic unit that embeds the sentence’s contextual information. In practice, during fine-tuning, each sentence of the training dataset was tokenized

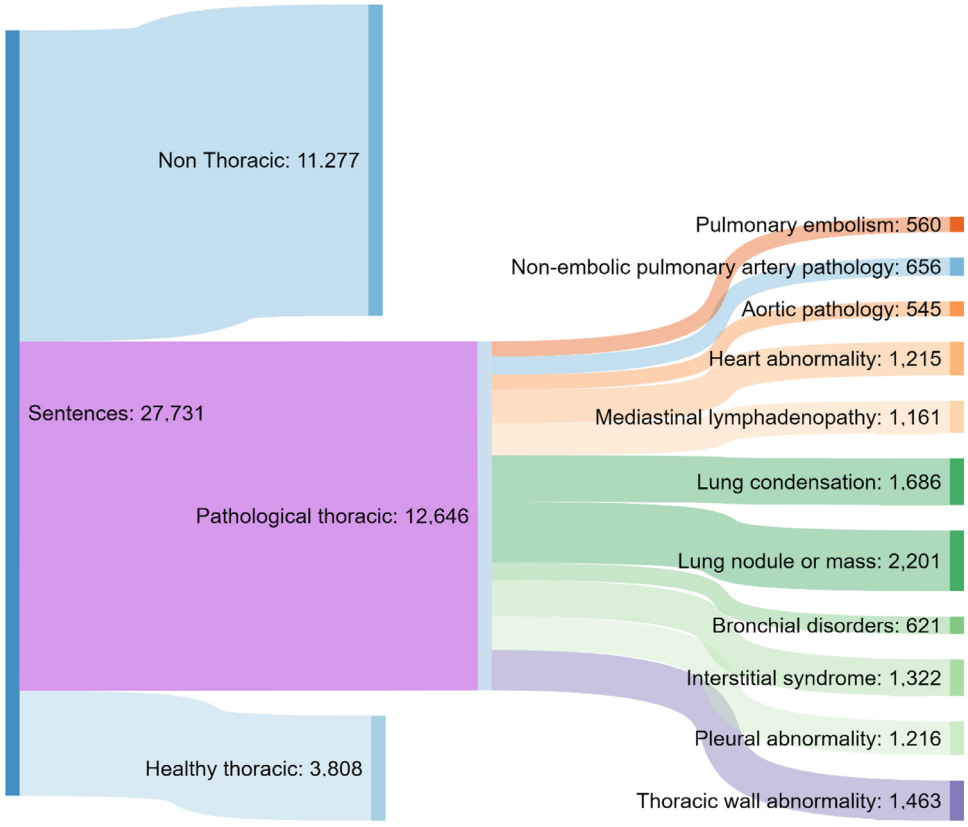

Fig. 1. Sankey Diagram of the 27731 labels used to train and validate the classifiers.

using the 'camembert-large' version of the Camembert tokenizer without taking labels into account. The original BERT [19] model uses a 15% probability of masking each token during model pre-training; we used a simplified version of this and assigned the same probability of each word being masked. At the second stage, we trained two classifiers in a supervised way based on the fine-tuned language model in the first stage using CamemBERTSequenceClassifier. The first, was trained on the 3-Classes sentences dataset. Then, the second classifier was trained on the 11 SubClasses sentences. Both classifiers were trained for 40 epochs with the same hyperparameters as above and reached convergence after 16 and 21 epochs respectively.

Both CamemBert Masked-Language-Model and CamemBert For Sequence Classification were implemented in PyTorch 1.7.0 and trained with a batch size of 32, 16 respectively and embedding size of 100 on a single NVIDIA Titan RTX GPU with 24 GiB of memory. We used an AdamW optimizer with learning rate 10^{-5} , betas (0.9, 0.999), and weight decay 10^{-2}. The network was trained with a standard cross-entropy (CE) loss.

For each of the two classifiers, accuracy, precision, recall and F1-score are reported. The standard metrics for evaluating sentence classification algorithms today, Precision and Recall. Precision and Recall can be formally defined as follows (where TP, FP, FN are the number of true positives, false positives, false negatives, respectively):

$$Precision = \frac{TP}{TP + FP}, \; Recall = \frac{TP}{TP + FN}$$

Informally, Precision is the fraction of all detected anomalies that are real anomalies, whereas, Recall is the fraction of all real anomalies that are successfully detected. In this sense, Precision and Recall are complementary, and this characterization proves useful when they are combined (e.g., using Fβ-Score, where β represents the relative importance of Recall to Precision).

2.3. Pulmonary embolism

Evaluation of the performance of the NLP model built in the previous section for PE detection was performed on the test dataset. We first identified all the chest CT examinations using a keyword analysis on the "Technical" part of the reports with the following keywords: "thor", "card", "coeur", "corps", "body", "pulm", "pneum", "TAVI", "RVAP", "tap", "coroscanner", "coro scanner", "coro-scanner". These stemmed keywords were chosen based on their representativeness in the technical sections of the examination covering the thoracic region. In order to select the examinations requested for pulmonary embolism detection, we then used a second keyword analysis on the "Clinical referral" part of the reports with the following stemmed keywords: "embo", "dimer", "'phlebi", "effet shunt" and "thromb vein profond'. The keywords given above correspond to the stemmed words for "thorax", "heart", "body", "thorax, abdomen and pelvis", "transcatheter aortic valve replacement (TAVR)", "coronary CT", "embolism", "d-dimer", "phlebitis", "shunt effect" and "thrombosis". Finally, all the selected reports, ie reports containing at least one keyword per set, were reviewed by two experienced radiologists who rated the reports as "Requested for pulmonary embolism detection" (yes or no) and "Positive for pulmonary embolism" (yes or no).

Following that, we performed our two stages classification workflow where we initially made a 3-Classes classifier on all the sentences from the "Findings" section of each report rated above as "Requested for pulmonary embolism detection". Consequently, sentences belonging to "Pathological chest" at first stage were then passed to 11-subClasses classifier. Reports were labeled "NLP Positive for pulmonary embolism" when one or more sentences within the same report were classified as "Pulmonary embolism".

Accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) were estimated for the prediction of the positivity of the report for pulmonary embolism detection in comparison with the human analysis. A qualitative analysis of the discrepancies between radiologists and the NLP method was also performed.

Table 1
French/English translated list of proposed words and their probability from the pretrained CamemBERT model (left column) and its fine-tuned version (right column) for the following masked sentence: "Absence de <mot masqué> du médiastin" / "No <masked word> of the mediastinum". One can note the strong improvement in relevance and certainty of the proposed word with the fine-tuned version of the pretrained CamemBERT model.

Before fine tuning		After fine tuning	
Words	**Probability**	**Words**	**Probability**
Traitement / Treatment	1.49 %	Déviation / Deviation	75.79%
Troubles / Trouble	1.18 %	Lésion / Lesion	7.49%
Maladie / Disease	1.17 %	Masse / Mass	3.33 %
Cancer / Cancer	0.87 %	Collection / Collection	4.45%

3. Results

Unsupervised fine-tuning of CamemBERT Masked Language pretrained model on validation dataset (50,350 sentences) resulted in an important improvement of the language model and drastic decrease of perplexity (4539.01 for CamemBERT versus 1.39 after fine tuning). An example of the results of the four most probable words proposed by the CamemBERT pretrained model and its fine-tuned version for a typical masked sentence is provided in Table 1.

The performance of classifiers is shown in Table 2. We report for both the 3-Classes and 11-SubClasses classifiers an accuracy, precision, recall and F1-score largely over 0.9. From the 94,559 scans from the test dataset, 4042 were requested for pulmonary embolism detection by the radiologists (4.27%) for which 641 (15.8%) were rated as "Positive for pulmonary embolism". Our NLP model rated 654 cases as positive which led to an accuracy, sensitivity, specificity, PPV and NPV of 99.13%, 98.28%, 99.29%, 96.33% and 99.68% respectively. Discrepancies between manual and NLP model rating were found in 35 reports of which 14 were related to an uncertainty on a PE, 7 to an unclear discord between acute and chronic PE, 3 to a partial regression of previously diagnosed PE, 4 to different diagnostic involving pulmonary arteries (e.g. tumor), 2 to the absence of PE evidence, which was missing in the description section but present in the conclusion and 5 to unaccountable errors.

4. Discussion

In this study, we trained a language-based model dedicated to CT report in French on a large corpus with a satisfying final perplexity score of 1.39. To put perplexity in context, we can in fact use two different approaches to evaluate language models: Extrinsic evaluation which involves evaluating the models by employing them in an actual task (such as sentence classification) and looking at their final loss and/or accuracy but this can be computationally expensive and slow as it requires training a full system. In contrast, Intrinsic evaluation involves finding some metric to evaluate the language model itself, not taking into account the specific tasks it's going to be used

Table 2
Results of the 3-category and the 11-category classifiers trained on a dataset of 27.731 sentences, for 5 hours using a NVIDIA Titan RTX, and validated on a dataset of ~1400 sentences.

Classifier	Accuracy	Precision	Recall	F1-score
3-category	0.988	0.979	0.989	0.984
11-category	0.984	0.986	0.985	0.985

for. Perplexity is an intrinsic evaluation metric for language models. Based on this model, classifiers of thoracic-wise abnormalities and normal sentences and sub-labelled pathological sentences reached F1-scores of 0.98. This allowed reporting a positivity yield rate of CT requested for PE detection of 16.2% with an accuracy of 99.1 % in comparison with the human-based assessment of the reports. This positivity rate is consistent with that recently published by Ben Cheikh et al. who reported a positivity rate of 15.8% in a large French cohort [20].

The use of NLP to analyze radiological reports is still relatively poorly explored, particularly for languages other than English [21]. Nevertheless, two European studies adopted NLP methods to detect pulmonary embolism from CT reports in French [10,21] using distinct technical approaches. Pham et al. reported equivalent performances, with a positive yield rate of 12.9 %, using a Maximum-Entropy classifier. In German, Weikert et al. tested different approaches to detect PE on CT reports and demonstrated a higher accuracy of CNN-based model with Word2Vec in comparison to Support Vector Machine (SVM) and Random Forest classification models applied on frequency-inverse document frequency representations. This finding is consistent with the results of other studies on NLP applied in different fields, showing the superiority of methods using word embeddings and particularly of those including the surrounding context of the words such as ULMFiT [22], OpenAI GPT [23], ELMo [24] and BERT [19,25–28]. Therefore, we decided to go with a contextual embeddings-based method in our study, and among the available methods, we chose a BERT-based model. Indeed, BERT contextual embeddings offer the biggest gains over non-contextual based methods in the area of capturing syntactic information and generally perform better when used in downstream learning tasks (e.g. sentence classification) [14]. BERT was also been shown superior to ELMo [26], Word2Vec [27] for emergency department chief complaints.

Another advantage of using a BERT-based approach is the transposability of our methodology in other languages since many pretrained models are available in open-source on several platforms (e.g. Hugging Face, https://huggingface.co) and their training is fully unsupervised. Over and above that, our language model in French can be used as a basis for other learning tasks and applications in France and other French speaking countries as French is the 5th most spoken language in the world, with 276.6 million speakers [29]. One should as well note that performing a two steps classification of the sentences with two classifiers to account for the imbalanced class problem related to the large differences between the number of “non-thorax” and “pathological thorax” subcategories [30,31].

To demonstrate the accuracy of our general chest CT reports classifier, we targeted the detection of PE given that it is a frequent indication for chest CT scan (up to 4.3 % of all the CT examinations performed in our institution). The relatively low positivity yield rate we found (15.8%) must be put into perspective given the potential severity of this pathology and the difficulties in its clinical detection. Nevertheless, it could also participate in a more general analysis of the appropriateness of CT for PE detection in low-risk clinical situations. Furthermore, a multi-label approach could allow other studies on the prevalence of different chest pathologies described in our labels such as aortic pathologies or interstitial lung diseases. We also point out that since our fine-tuned model included all the CT examinations performed at our institution, it can be employed to develop other classifiers for neuro or abdominal CT reports for instance.

We demonstrated throughout the sections above the effectiveness of the NLP model in classifying PE. However, two main challenges must be underlined and remain unexplored by our study: reports uncertainty and multi-label classification settings. The former, is a serious issue in daily radiological practice leading to misunderstandings and diagnostic errors [32–34]. Thus, the medical diagnoses often appear to be very subjective to human readers (experienced radiologists). On that account, we investigated misclassified reports and reported that 60 % of them are related to uncertainty (e.g: “doubt on a small distal endovascular defect in the left anteromedial basal segment which might be related to cardiac motion”). Various attempts have been made to try to quantify medical uncertainty in reports with NLP [35] but still difficult to handle [36].

The latter, related to mono-label versus multi-label settings where classification assigns to each sample a set of target labels. This can be thought of as predicting properties of a data-point that are not mutually exclusive, such as different abnormalities that are relevant for a CT report. A sentence might be about “Heart abnormality” and “Pleural abnormality” at the same time or none of these , e.g : “presence of a pericardial and bilateral pleural effusion” would belong to both “Heart abnormality” and “Pleural abnormality”. In spite of that, we consider mono-label settings which increase the robustness of the decision of the model for each label but may be misleading in case of a sentence including several labels. Nevertheless, we believe that the impact of this problem seems to be relatively low as we chose to focus on the description part of the report where each finding is generally described separately.

Finally, a limitation of our study is that we did not compare our transformer-based approach with other NLP methods, including simpler methods such as the one based on NegEX or FastContext [37]. This comparison would be interesting but would require an implementation of many other NLP methods, which is out of the scope of our study.

5. Conclusion

In conclusion, BERT-based NLP model allows achieving high performances for sentence classification on French chest CT reports. The methodology we developed to measure the positivity yield rate of PE could be extended to other clinical situations and participate in the evaluation of the appropriateness of CT in these different settings.

Ethical_statements

Human and animal rights

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans as well as in accordance with the EU Directive 2010/63/EU for animal experiments.

Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s).

Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

Conceptualization L. Boussel Data curation L. Boussel ; F. Talbot ; A. Richard ; S. Gouttard ; A. Mansuy Formal Analysis L. Boussel ; F. Talbot ; A. Richard ; S. Gouttard ; A. Mansuy Methodology L. Boussel Project administration L. Boussel Resources L. Boussel ; F. Talbot ; A. Richard ; S. Gouttard ; A. Mansuy Software L. Boussel ; F. Talbot ; A.

Richard ; S. Gouttard ; A. Mansuy Supervision P. Douek Validation L. Boussel Writing – original draft E. Jupin-Delevaux Writing – review & editing A. Djahnine ; L. Boussel

Declaration of Competing Interest

The authors declare that they have no known competing financial or personal relationships that could be viewed as influencing the work reported in this paper.

References

[1] Kirsch J, Brown RKJ, Henry TS, et al. ACR Appropriateness Criteria® Acute Chest Pain—Suspected Pulmonary Embolism. J Am Coll Radiol 2017;14:S2–S12. doi: 10.1016/j.jacr.2017.02.027.

[2] Whitehead MT, Cardenas AM, Corey AS, et al. ACR appropriateness criteria® headache. J Am Coll Radiol 2019;16:S364–77. doi: 10.1016/j.jacr.2019.05.030.

[3] European Society of Radiology (ESR). Methodology for ESR iGuide content. Insights Imaging 2019;10:32. doi: 10.1186/s13244-019-0720-z.

[4] Goldzweig CL, Orshansky G, Paige NM, et al. Electronic health record–based interventions for improving appropriate diagnostic imaging: a systematic review and meta-analysis. Ann Intern Med 2015;162:557. doi: 10.7326/M14-2600.

[5] Blackmore CC, Mecklenburg RS, Kaplan GS. Effectiveness of clinical decision support in controlling inappropriate imaging. J Am Coll Radiol 2011;8:19–25. doi: 10.1016/j.jacr.2010.07.009.

[6] Ganeshan D, Duong P-AT, Probyn L, et al. Structured reporting in radiology. Acad Radiol 2018;25:66–73. doi: 10.1016/j.acra.2017.08.005.

[7] Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016;279:329–43. doi: 10.1148/radiol.16142770.

[8] Chen P-H, Zafar H, Galperin-Aizenberg M, Cook T. Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. J Digit Imaging 2018;31:178–84. doi: 10.1007/s10278-017-0027-x.

[9] Li Y, Dai Y, Deng L, et al. Computer-aided detection for the automated evaluation of pulmonary embolism. THC 2017;25:135–42. doi: 10.3233/THC-171315.

[10] Weikert T, Nesic I, Cyriac J, et al. Towards automated generation of curated datasets in radiology: application of natural language processing to unstructured reports exemplified on CT for pulmonary embolism. Eur J Radiol 2020;125:108862. doi: 10.1016/j.ejrad.2020.108862.

[11] Kang SK, Garry K, Chung R, et al. Natural language processing for identification of incidental pulmonary nodules in radiology reports. J Am Coll Radiol 2019;16:1587–94. doi: 10.1016/j.jacr.2019.04.026.

[12] Katzman BD, van der Pol CB, Soyer P, Patlas MN. Artificial intelligence in emergency radiology: a review of applications and possibilities. Diagnost Intervent Imag 2023;104:6–10. doi: 10.1016/j.diii.2022.07.005.

[13] Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs].

[14] Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. J Am Med Inform Assoc 2020;27:1935–42. doi: 10.1093/jamia/ocaa189.

[15] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019 btz682. doi: 10.1093/bioinformatics/btz682.

[16] Li F, Jin Y, Liu W, et al. Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: an empirical study. JMIR Med Inform 2019;7:e14830. doi: 10.2196/14830.

[17] Pota M, Ventura M, Catelli R, Esposito M. An effective BERT-based pipeline for twitter sentiment analysis: a case study in Italian. Sensors 2020;21:133. doi: 10.3390/s21010133.

[18] Martin L, Muller B, Suárez PJO, et al. CamemBERT: a tasty French language model. In: Proceedings of the 58th annual meeting of the association for computational linguistics; 2020. p. 7203–19. doi: 10.18653/v1/2020.acl-main.645.

[19] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–86.

[20] Cheikh AB, Gorincour G, Nivet H, et al. How artificial intelligence improves radiological interpretation in suspected pulmonary embolism. Eur Radiol 2022. doi: 10.1007/s00330-022-08645-2.

[21] Casey A, Davidson E, Poon M, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak 2021;21:179. doi: 10.1186/s12911-021-01533-7.

[22] Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. arXiv:180106146 [cs, stat].

[23] Radford et al. Improving language understanding by generative pre-training. Accessed 5 Nov 2021

[24] Peters ME, Neumann M, Iyyer M, et al (2018) Deep contextualized word representations. arXiv:180205365 [cs].

[25] Tenney I, Xia P, Chen B, et al (2019) What do you learn from context? Probing for sentence structure in contextualized word representations. arXiv:190506316 [cs]

[26] Chang D, Hong WS, Taylor RA. Generating contextual embeddings for emergency department chief complaints. JAMIA Open 2020;3:160–6. doi: 10.1093/jamiaopen/ooaa022.

[27] Saha B, Lisboa S, Ghosh S. Understanding patient complaint characteristics using contextual clinical BERT embeddings. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC); 2020. p. 5847–50.

[28] Gao Z, Feng A, Song X, Wu X. Target-dependent sentiment classification with BERT. IEEE Access 2019;7:154290–9. doi: 10.1109/ACCESS.2019.2946594.

[29] (2021) https://www.ethnologue.com/family/17-3560. In: Ethnologue. https://www.ethnologue.com/family/17-3560. Accessed 15 Sep 2021

[30] Zahera HM (2019) Fine-tuned BERT model for multi-label tweets classification. In: TREC

[31] Cai L, Song Y, Liu T, Zhang K. A hybrid BERT model that incorporates label semantics via adjustive attention for multi-label text classification. IEEE Access 2020;8:152183–92. doi: 10.1109/ACCESS.2020.3017382.

[32] Bruno MA. 256 Shades of gray: uncertainty and diagnostic error in radiology. Diagnosis (Berl) 2017;4:149–57. doi: 10.1515/dx-2017-0006.

[33] Abujudeh HH, Boland GW, Kaewlai R, et al. Abdominal and pelvic computed tomography (CT) interpretation: discrepancy rates among experienced radiologists. Eur Radiol 2010;20:1952–7. doi: 10.1007/s00330-010-1763-1.

[34] Bruno MA, Petscavage-Thomas J, Abujudeh HH. Communicating uncertainty in the radiology report. Am J Roentgenol 2017;209:1006–8. doi: 10.2214/AJR.17.18271.

[35] Reiner BI. Quantitative analysis of uncertainty in medical reporting: creating a standardized and objective methodology. J Digit Imaging 2018;31:145–9. doi: 10.1007/s10278-017-0041-z.

[36] Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning. ACM Comput Surv 2021;54 115:1-115:35. doi: 10.1145/3457607.

[37] Mirzapour M, Abdaoui A, Tchechmedjiev A, et al. French FastContext: a publicly accessible system for detecting negation, temporality and experiencer in French clinical notes. J Biomed Inform 2021;117:103733. doi: 10.1016/j.jbi.2021.103733.