Automatic analysis of negation cues and scopes for medical texts in French using language models

Salim Sadoune, Antoine Richard, François Talbot, Thomas Guyet, Loïc Boussel, Hugues Berry

Automatic analysis of negation cues and scopes for medical texts in French using language models

S. Sadoune[a], A. Richard[b], F. Talbot[b], T. Guyet[a,d], L. Boussel[b] and H. Berry[a,d,*]

[a]*Inria, Lyon Research Center, F-69603, Villeurbanne, France*
[b]*DSN Bron, Hospices Civils de Lyon, F-69672, Bron, France*
[c]*CREATIS UMR 5220, INSA-Lyon, Université Claude Bernard Lyon 1, CNRS, Inserm, Villeurbanne, F-69621, Lyon, France*
[d]*AIstroSight, Inria, Université Claude Bernard Lyon 1, Hospices Civils de Lyon, Villeurbanne, F-69603, France*

ARTICLE INFO

Keywords:
Negation Analysis
Language Models
Medical Reports
Token classification task
Transformers

ABSTRACT

Objective
Correct automatic analysis of a medical report requires the identification of negations and their scopes. Since most of available training data comes from medical texts in English, it usually takes additional work to apply to non-English languages. Here, we introduce a supervised learning method for automatically identifying and determining the scopes and negation cues in French medical reports using language models based on BERT.
Methods
Using a new private corpus of French-language chest CT scan reports with consistent annotation, we first fine-tuned five available transformer models on the negation cue and scope identification task. Subsequently, we extended the methodology by modifying the optimal model to encompass a wider range of clinical notes and reports (not limited to radiology reports) and more heterogeneous annotations. Lastly, we tested the generated model on its initial mask-filling task to ensure there is no catastrophic forgetting.
Results
On a corpus of thoracic CT scan reports annotated by four annotators within our team, our method reaches a F1-score of 99.4% for cue detection and 94.5% for scope detection, thus equaling or improving state-of-the art performance. On more generic biomedical reports, annotated with more heterogeneous rules, the quality of the automatic analysis of course decreases, but our best-of-the class model still delivers very good performance, with F1-scores of 98.2% (cue detection), and 90.9% (scope detection). Moreover, we show that fine-tuning the original model for the negation identification task preserves or even improves its performance on its initial fill-mask task, depending on the lemmatization.
Conclusion
Considering the performance of our fine-tuned model for the detection of negation cues and scopes in medical reports in French and its robustness with respect to the diversity of the annotation rules and the type of biomedical data, we conclude that it is suited for utilization in a real-life clinical context.

1. Introduction

Radiology reports are textual minutes recorded by radiologists to summarize their interpretation of a medical imaging examination. These medical reports have gained increasing interest as it has been realized that they represent a rich and easily available source of information for the extraction of clinical conditions or assistance to diagnosis [39, 3, 41]. For instance, they can be used to automatically identify patients suffering from pulmonary embolism or bronchial disorders or other labels of interest for the physician [16]. Because of their narrative form, such an automatic analysis demands dedicated natural language processing (NLP) approaches to extract the information of interest in a structured way from the raw text. Most of these approaches have focused on named entity recognition tasks, i.e. the classification of words into predefined medical classes [42, 19, 27]. However, less attention has been paid to context analysis, i.e. the analysis of the surrounding of a word or a sentence, that helps clarify the content, such as the identification of negations and their scopes [6, 32, 40].

*Corresponding author
hugues.berry@inria.fr (H. Berry)
ORCID(s): 0000-0001-8677-8910 (A. Richard); 0000-0002-4909-5843 (T. Guyet); 0000-0001-9053-8127 (L. Boussel); 0000-0003-3470-683X (H. Berry)

Automatic identification of a negation and of its scope, i.e., the section of the sentence to which it applies, is however critical to a correct interpretation of a medical report. For instance, in the sentence: “No proximal pulmonary embolism”, the presence of the negation indicator “No” at the beginning signals an absence of thoracic pathology where a simplistic named entity recognition algorithm, ignoring the negation indicator to focus on the medical condition only, would conclude in favor of a pathology. Beyond this obvious example, the negation context in real-world records is frequently more complex. Consider for instance the following example, translated from a medical report in french: “The patient is a 60-year-old man with breathing difficulties, non diabetic, appetite - good, no chest pain, no weight loss nor episodes of stomach pain, hypertension absent”. In this real-case example, the negation is distributed across several segments of the sentence to offer a comprehensive depiction of the patient’s medical condition. The main challenge is therefore to correctly estimate what part of the sentence is concerned by every negation cues. For instance, in “non diabetic, appetite - good”, the scope of “non” should be restricted to “diabetic”and not attached to “appetite”. Also, at the end of the sentence (“hypertension absent”), the negation cue is located after its scope, not before it as is usually the case.

Another difficulty of negation cue and scope identification comes from the necessity for the developed algorithms to be specifically defined for or trained to a target language, since the ways by which negation and its scope are built, are often language-specific. With most of the available training data extracted from reports in English, the application to non-English languages is less developed and often necessitates dedicated efforts [10, 28, 37, 15].

In the present work, our first objective was to propose a recognition pipeline for negation based on language models specifically for french radiology reports with coherent annotation. We then tested the ability to generalize our approach by applying it to various clinical notes and reports (beyond radiology reports) and with more heterogeneous annotations. Finally, we checked the absence of catastrophic forgetting, by comparing the performance of the model fine-tuned for negation detection with that of its original version before fine-tuning, on the usual fill-mask task.

2. Related work

Negation detection in text has mostly been using rule-based systems [14] and supervised machine learning-based approaches [7].

2.1. Rule-based systems

A rule-base system, a.k.a. expert system, denotes a computer program that utilizes a predefined set of rules to reason about facts. In rule-based systems for NLP, the rules are established by experts. These systems remain extensively utilized in the biomedical domain for extracting information from medical reports, which includes the identification of negated information [31]. Based on regular expressions, NegEx [4] was the first rule-based proposal allowing for negation detection in clinical texts in English and has been adapted to non-English languages, including French [10]. The performance of the initial rule-based systems left room for improvement with F1-scores only slightly larger than 80% for the detection of negation cues [4] and even lower for negation scopes [2]. More recent proposals improved the performance mostly by incorporating some of the semantics of the sentences, yielding e.g., F1-scores ranging from 93% to 96% for cue detection [23, 29] on biomedical text corpora.

In spite of their simplicity and efficiency, rule-based systems inherently come with limitations. Clinical notes and reports use precise terminology but may contain spelling errors, short sentences with a number of abbreviations (example 1 bellow), and acronyms (example 2 below). Additionally, some terms used in French as negation indicators can be polysemous (see example 3 below). Consider for instance the following sentences, where abbreviation, acronym, or polysemous term are shown in brackets:

1. Leur cytoplasme abondant, granuleux et éosinophile était fortement coloré par le [PAS](Periodic-Acid-Schiff).
2. Absence d’[EP].
3. Debout, pieds parallèles, bras le long du corps, ouvrez le pied droit sur le côté et avancez d’un bon [pas] avec la jambe gauche.
4. les études d’interactions [n’] ont été réalisées que chez l’adulte.

In the first sentence, that was extracted from a biomedical corpus in French [8], the negation term “pas” appears as an abbreviation for the Periodic Acid-Schiff (PAS) staining[1]. This complicates the task of an expert system in

[1] https://www.biovalley.fr/achat/cat-coloration-pas-periodic-acid-schiff-5391.html

managing the context. The second example shows a sentence from a CT scan report where the radiologist used the acronym "EP" (for *Pulmonary Embolism*) to specify the type of thoracic pathology. The information of this acronym must be explicitly given to a rule-based system for it to identify pulmonary embolism/EP as a negated entity. The third sentence was extracted from a scientific article. Here, the name "pas" (step) is a pseudo-indicator, that does not represent a negation in the sentence. In the last sentence, the use of the negation term "n'" with the conjunction "que" indicates a restriction, not a negation. An expert system which rule base would not encompass all this information would erroneously identify the pseudo-indicator as a negative term and transform this affirmative sentence into a negative one.

2.2. Machine learning

In machine learning-based approaches, negation detection is considered a sequence labeling task [25], which assigns labels for negation indicators and their scope. The early availability of corpora of biomedical texts in English, annotated for negation cues and scopes, such as BioScope [33], has fostered the exploitation of a range of machine or deep learning approaches including support vector machines (SVM) [38], conditional random fields (CRF) [21], convolutionnal neural networks (CNN) [30], Bidirectional Long Short-Term Memory (Bi-LSTM) [12, 34] or transformers [17]. On generic biomedical texts, like BioScope, these approaches exhibit very good performance, with F1-scores that are frequently well above 96% for cue detection [24, 1]. The detection of negation scope is here also a more difficult task. For most of the proposed approaches, the reported F1-scores for scope detection are closer to 90% [24, 30, 12, 21] though more recent approaches outperformed them by 3 to 6 F1-score at the price of a decreased performance on cue detection, though [17].

Beyond English, several studies have been published on the detection of negations within biomedical corpora in other languages, including Spanish [28], and Dutch [37]. Regarding French, to our knowledge, the first machine learning-based negation detection system in biomedical texts written in French was introduced in the work of Daloux et al. [8], where the use of a number of families of machine learning approaches was explored (CRF, BiLSTM, Bi-LSTM). With CRF and BiLSTM, for instance, their approach reached a F-measure of 97.2 % for the cue and 90.8 % for the scopes on coherently annotated biomedical texts [8].

As with any machine learning approach, the accessibility of a large and annotated learning set is crucial. The availability of large amounts of annotated medical reports in French is still a challenge because they can contain sensitive personal data. For this reason, the availability of large corpora of medical reports is usually restricted to the hospital that produced them. A handful of anonymous datasets of french biomedical reports can however be used to train algorithms, although they usually do not consist in radiology reports. Daloux et al. [8] assembled two annotated biomedical corpora with negation information: the ESSAI clinical corpus contains clinical trial protocols in French that have been collected from the registry of the National Cancer Institute [9], and the CAS corpus [13] in French comprises clinical cases as published in scientific, legal, or educational literature, describing clinical situations for real de-identified or fake patients. The two corpora have been annotated at three levels. First, a morpho-syntactic tagging (or PoS – Part of Speech tagging), i.e. the annotation of the relationship between word structure and sentence construction, was performed automatically using tagex[2]. Then, annotation of negation cue and negation scope was carried out manually by the authors. Table 1 illustrates the annotation available in the ESSAI corpus with information provided for each token of a short sentence. PoS-tag indicates the morpho-syntactic tagging (PoS-tag), `M-neg` indicates the negation cues with two new types of tags (`B-cue-neg` and `I-cue-neg`) and `P-neg` is related to the negation scope. Note that the absence of tag in the original dataset (illustrated by "-" in the table) have been replaced by `O` (other) in our dataset. Both corpora are freely available on request for research purposes[3] under a CC BY-NC-SA 4.0 DEED license. A statistical summary of these two biomedical corpora is given in tables 1 and 2 in the Appendix.

3. Materials and methods

In this section, we present the materials and methods we employed in our study. We start by presenting a new dataset, RADIO, which has been created to supplement the two existing datasets mentioned above. Then, we present the methodology to fine-tune and evaluate state-of-the-art token classifiers for negation detection.

[2] https://allgo.inria.fr/app/tagex

[3] https://clementdalloux.fr/?page_id=28

Token	PoS-tag	M–neg	P–neg
Cette	PRO:DEM	-	-
étude	NOM	-	-
ne	ADV	B-cue_neg	-
modifiera	VER:futu	-	B-scope_neg
pas	ADV	I-cue_neg	-
la	DET:ART	-	I-scope_neg
prise	NOM	-	I-scope_neg
en	PRP	-	I-scope_neg
charge	NOM	-	I-scope_neg
thérapeutique	ADJ	-	I-scope_neg
.	SENT	-	-

Table 1
A sentence extracted from the ESSAI corpora and its annotations. The "Token" column lists the consecutive tokens of the sentence. For each token, the "PoS-tag" column shows its morpho-syntactic tag: PRO:DEM (demonstrative pronoun), NOM (noun), ADV (adverb), VER:futur (verb future tense), DET:ART (article), PRP (preposition), ADJ (adjective) or SENT (sentence tag). Negation annotation is given in the two last columns for negation cue (M-neg) and scope (P-neg): B-cue_neg (beginning negation cue), I-cue_neg (internal negation cue), B-scope_neg (beginning of the negation scope) or I-scope_neg (token within the negation scope).

3.1. RADIO, a corpus of radiology reports in French

While the availability of open resources like the ESSAI and CAS corpora is a strong asset for the development of negation annotation systems for generic biomedical texts, it does not allow to build a system for radiology reports specifically, which is the primary objective of the present work. We therefore built a new corpus of radiology reports in french, based on clinical data provided by the Radiology Department of Lyon University Hospital (Hospices Civils de Lyon, HCL), France's second university hospital centre[4].

To build a radiology-specific corpus, we gathered 10,798 non-duplicate anonymized single sentences extracted from thoracic CT scan reports provided by the Radiology Department of HCL, together with their tagging with 23 thoracic pathology classes (pleural anomaly, cardiac anomaly, ...). We refer to this corpus as RADIO. Among this set, 2,321 sentences actually contains at least one negation cue and were specifically tagged for negation cue and scope. The other sentences have been discarded from the training dataset. In total, RADIO contains 17 different negation indices. The most frequent negation marker is the term "pas" (not), with 707 occurrences, while the least frequent marker consists of terms such as "absent", "jamais", "aucune ... n'", "ne ... jamais", with only one occurrence each. Tables 1 and 2 in the Appendix show the main statistics about this corpus.

For the tagging of negation cues and scopes, the RADIO corpus was manually annotated by a group of 4 annotators according to the following process and rules. From a list of negation indices encountered in French taken from [8], we annotated manually each negation cue in each negated sentence, as well as its scope (i.e., all the tokens in the sentence that are affected by the negation cues). Importantly, we distinguished total from partial or exceptive negations. Partial negations concern only one element of the sentence (sentence 1 below), whereas total negations negate the whole idea expressed by the sentence (sentence 2 below). Total negations mostly use specific negation adverbs like "n'(ne) ... pas/point". Exceptive negations refers to cases where negation indicators does not correspond to a negative sentence, but rather to some form of restriction (sentence 3 below). They are usually expressed using "n'/ne ... que". Note that the restriction expressed by an exceptive negation "ne ... que" can be negated by the insertion of the adverb "pas" between "ne" and "que" (sentence 4 below).

A first set of annotations was made by a single annotator on the 2,321 sentences that contain a negation, taking into account that exceptive negations can refer to restriction, not negation. These annotation proposals were then forwarded to a group of three distinct annotators for re-annotation. Disagreements were observed among annotators, particularly regarding the endpoint of the scope. To resolve a disagreement on the scope, we searched for the presence in the sentence of a negated stopping point, based on a list of stopping point tokens that are known to change the meaning of the information, such as "mais", "sous réserve", "en raison", or "au niveau" ("*but*", "*subject to*", "*due to*", or "*at the level of*" in English). An example where the scope stops at token "au niveau" is shown in sentence 5 below. In the case

[4] https://teamhcl.chu-lyon.fr/about-us

of negative sentences that did not have such stop tokens, a linguistic check was performed to resolve disagreements at the scope endpoint, making sure to incorporate additional information, such as adjectives, to enhance the description of the extracted negative information (see e.g., sentence 6 below).

Below is a list of sentences to illustrate the result of our negation cue and scope annotation. Annotation cues are in brackets and their scope is underlined. It can happen that no decisive argument exists in favor of the inclusion of a token in the scope, or in favor of its exclusion. In this case, we highlight the corresponding tokens with a dotted underline:

1. Discrets troubles ventilatoires des deux bases [non] spécifiques.
2. La dérivation cardiopulmonaire[n']est[pas]opacifiée.
3. À noter que il [n'] est inséré sur le mur en postérieur [que] par un pont filiforme.
4. L' artère marginale [n'] est [pas] décelable [que] en aval du pontage.
5. L'[absence d'] anomalie au niveau de l' aorte thoracique ascendante.
6. [Pas] d'anomalie du pancréas , et des surrénales.

3.2. Negation detection in french radiology reports

In terms of learning systems, we selected transformer-based models from the literature, fine-tuned them for negation detection and benchmarked their performance.

We selected five NLP transformer models:

- CamemBERT-base [22] is a pre-trained model for the French language, based on the Roberta architecture [11], and trained on 138GB of text data from the OSCAR dataset.
- Camembert-bio [35] is a state-of-the-art french biomedical language model built using continual-pretraining from CamemBert-base and pre-trained on 413 million tokens of biomedical corpus.
- The DrBERT model [20] is also based on the RoBERTa model architecture, but was trained from scratch on a 7GB corpus of French medical text data.
- RadBERT [41] is a transformer-based language model specifically adapted for radiology by pre-training with millions of radiology reports in English. Here, we used its variant RadBERT-RoBERTa-4m[5].
- FrRadBERT is a fine-tuning of RadBERT that we produced by fine-tuning RadBERT on the RADIO corpus (task: Masked Language Modeling, batch size 16, Adam optimizer [18] with a training rate of 5×10^{-5}).

Each of the 5 models above were fine-tuned for negation detection with the HuggingFace Transformers library. To fine-tune the models, one first needs to convert each token in the input sentence into its unique identifier that will be used as input by the model. To this aim, we tokenized the inputs by associating each model with its own tokenizer. However, the models in French are based on a tokenizer for sentences with a characteristic independent of accents. Therefore, the tokenizer of french models has been expanded with a list of words undergoing elision ("n'", "l'", "s'", ...) to effectively handle white spaces during the tokenization of apostrophes. We splitted out-of-vocabulary words into sub-tokens, assigning a label to the first sub-word and ignoring others (label `IGN`). The offsets between the annotations that result from this process were corrected by a function that allows each token to be aligned and to be associated with its label. With the data thus prepared, aligned and tokenized, the language model based on the transformer to benchmark was trained and evaluated. Each token classifier was trained to predict the tags related to negation in a supervised manner using the Adam optimizer with a learning rate of 5×10^{-5} and a batch size of 16.

Each model was trained using cross-validation with 10 folds, resulting in an 90% training set and a 10% validation. For each fold, the model was trained for 25 epochs on a single V100 GPU architecture.

Evaluation was made in two phases. In the first phase, we assessed the performance of each model by evaluating its ability to predict the complete class, i.e., to correctly identify both the negation cue and the entire sequence of tokens that contains the scope. We computed the mean and standard deviations over the cross-validation for three metrics: precision (P), recall (R), and F1-score ($F1$).

[5] https://huggingface.co/zzxslp/RadBERT-RoBERTa-4m

A second evaluation phase was then performed in order to deepen our understanding of the limitations of the models in negation recognition via a detailed analysis of prediction errors. We adopted an approach inspired by [36], which entails scrutinizing validation sentences with the highest loss by computing the loss per token. Throughout this process, the cross-entropy loss is calculated between the logits and each token to find the most likely label.

3.3. Generalizing to biomedical records with less coherent annotation

The context of the first experiments is more favorable than what would be expected from a practical implementation in the clinical routine. First, the negative sentences were annotated using highly consistent rules for the annotation (section 3.1). In a clinical context, the rules for negation identification, especially for the scope length, would probably be much more inconsistent, from an expert to the other. Moreover, the dataset used to fine-tune the model is limited to thoracic CT scan reports, a scope that is probably too limited to be really of interest in a clinical context. To address these questions, We enriched our learning set with the two biomedical report corpora: ESSAI and CAS, introduced in Section 2.2. Those corpora are not restricted to radiology reports but mostly feature texts from a broad biomedical context. Furthermore, they were annotated by different experts, not co-authors of the present study who followed annotation rules less uniform than the ones we imposed for the RADIO corpus.

Following the procedure of the previous section, we retained the same five transformer models. Each model can be fine-tuned to each of the three datasets and their fusions, resulting in a too large number of combinations. We therefore selected four illustrative combinations.

Scenario 1: This is the most straightforward scenario: we took the CamemBERT-bio model that was fine-tuned on the negation task with the RADIO corpus in section 3.1 above, and tested it on a merging of the biomedical corpora ESSAI+CAS.

Scenario 2: This second scenario implements the reverse procedure of the first one: we fine-tuned three models, CamemBERT-base, CamemBERT-bio and DrBERT on the merged ESSAI+CAS corpora, then evaluated the resulting refined model on the RADIO corpus.

Scenario 3: As a variant of Scenario 2, we fine-tuned CamemBERT-base, CamemBERT-bio and DrBERT on the fusion of the RADIO and ESSAI corpora, before assessing its performance on the CAS corpus.

Scenario 4: This final scenario was elaborated to exploit a corpus maximizing the quantity of information and annotation criteria. We built a consolidated dataset by merging our three annotated corpora above (RADIO, CAS, and ESSAI). We also added 109 negative sentences extracted from the QUAERO medical corpus [26] and 12 additional sentences extracted from French scientific articles containing complex cases of exceptive negation. The resulting corpus, that we refer to as the Negation Large Medical French Corpus (NLMFC), contains 21, 953 sentences, including 4, 244 negative ones. Three language models (CamemBERT-base, CamemBERT-bio and DrBERT) were fine-tuned and evaluated on this corpus, using 10-fold cross-validation.

3.4. Over-fitting Audit

Finally, we complemented our experiments with an evaluation of the potential over-fitting of the fine-tuned model, also known as the catastrophic forgetting effect [43]. Indeed, fine-tuning a model to detect negations can worsen its original capabilities on other tasks. We therefore assessed whether a model fine-tuned on an auxiliary task (negation annotation) still performs accurately on the original task of the model (fill-mask).

For this experiment, we compared the original and fine-tuned versions of the CamemBERT-bio model. The fine-tuned version was that of scenario 1 above, i.e., CamemBERT-bio fine-tuned on the entire RADIO dataset. Since CamemBERT-bio was originally trained on a fill-mask task, we compared the oringial CamemBERT-bio and the version we fine-tuned on fill-mask task based on perplexity [5]. Perplexity measures how well a language model predicts a text sample. It is calculated as the average number of bits per word a model needs to represent the sample. The lower the value, the better the model. We calculated the mean perplexity of the masked fill-in-the-blank task on lemmatized or lemmatized sentences from 11, 037 samples from the ESSAI+CAS dataset.

4. Results

4.1. Accuracy of fine-tuned token classifiers for negation detection

This section presents the results we obtained for the detection of negation cues and scopes on the RADIO corpus of chest CT scan reports in French presented in Section 3.1.

	Cue detection			Scope detection		
Models	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
CamemBERT-base	99.18 ± 0.29	99.31 ± 0.30	99.25 ± 0.25	93.36 ± 0.96	94.50 ± 0.81	94.17 ± 0.83
CamemBERT-bio	$\mathbf{99.35 \pm 0.24}$	$\mathbf{99.37 \pm 0.31}$	$\mathbf{99.36 \pm 0.25}$	$\mathbf{94.19 \pm 0.94}$	$\mathbf{94.80 \pm 0.84}$	$\mathbf{94.49 \pm 0.80}$
DrBERT-7GB	99.04 ± 0.35	99.34 ± 0.24	99.19 ± 0.25	91.38 ± 1.13	93.13 ± 0.76	92.24 ± 0.87
RadBERT-4m-fine-tuning	99.04 ± 0.31	99.30 ± 0.23	99.17 ± 0.19	91.17 ± 1.44	91.95 ± 0.88	91.56 ± 0.96
FrRadBERT-4m-fine-tuning	99.18 ± 0.25	99.33 ± 0.27	99.26 ± 0.25	92.19 ± 0.98	93.29 ± 1.012	92.74 ± 0.94

Table 2
Performance of the five tested models on the negation cue detection and negation scope detection tasks on the RADIO corpus of french chest CT scan reports. Results are given as average±standard-deviation of Precision, Recall, and F1-score computed over the 10-fold cross-validation. Underlined numbers locate a statistically significant difference with the best model, shown in bold, using a *t*-test with a significance level of 5%.

Table 2 presents the results obtained on the validation set by comparing the scores obtained with our five language models. All the five models successfully detected almost all negation cues as indicated by their F1-scores, that were all larger than 99% for cue detection. All models performed equally well on this task, with no statistical difference between their performance (*t*-test, significance level 5%).

CamemBERT-base and CamemBERT-bio exhibited the best performances for scope detection, with no significant difference between them. This suggests that the additional fine-tuning of CamemBERT-bio on biomedical data does not help much for our negation labeling task. This result is not really surprising, since the way negation is expressed in non-medical texts in french is not expected to radically differ from the way it is expressed in medical texts. Scope detection remains a more complex task than cue detection. The F1-scores for scope detection were indeed markedly reduced compared to the cue detection task, by approximately 5 for the two best-of-the-class models (to around 94%) and by circa 7 for the others (to around 92%).

The worse performance on scope detection was obtained with the RadBERT-4m-fine-tuning model, a model trained on English radiology reports. This indicates that making accurate predictions about the scope requires a solid understanding of linguistics. In agreement with this interpretation, we also observed a slight improvement in performance after adaptation of the RadBERT model to French data (FrRadBERT-4m-fine-tuning).

To deepen our understanding of the model performances, we ran a thorough error analysis of the previous results, for the best-performance model, i.e. CamemBERT-bio. We selected the best model from those obtained during the 10 cross-validation folds. Our error analysis involves predicting all the sentences in the validation set and calculating the token-wise loss for each sequence, we use cross-entropy loss to measure these values. Table 3 displays an example of such a predicted sentence from the validation set, illustrating the loss value associated with each input token, along with the predicted label and the corresponding true label.

In table 4 , we examined the validation examples with the highest loss, calculated the loss per token in the sequence, then grouped it by input tokens and aggregated the losses for each token with the count, the mean, and the sum. We observed that:

- The token “de” represent the highest total loss, which is not surprising since it is also the most common token in the list of sentences. However, its mean loss is much lower than that of the other tokens in the list. This means that the model has no difficulty in classifying it, as well as the stopwords “en”, “un”, “du”.
- The tokens “tasse”, “plateau” and “spontané” represent the highest mean loss values in the sentences in the validation set. These tokens are generally present in sentences (sentence 3, for example) where the model often confuses the end of the scope with its inside.

Since each input token is associated to a label by the model, we can group the data validation sentences by label and aggregate the losses for each label class by computing their count and the mean of the losses. The results are displayed in Table 5. The inside of the negation scope (`I-scope_neg`) by far displays the highest average loss (0.32), which is unsurprising considering it is also the most common entity (733 tokens) after the majority entity `O`, for “Others” (14, 443 tokens). This indicates that the main challenge for the model consists in accurately identifying the inside of the scope. This is consistent with the overall results in Table 2 showing a slight drop of F1-score from cue detection

True labels	Loss	Predicted label	Token
B-cue_neg	0.0	B-cue_neg	_non
B-scope_neg	0.0	B-scope_neg	_retenue
O	1.68	I-scope_neg	_comme
O	1.84	I-scope_neg	_pathologique

Table 3
A portion of a predicted sentence from the data validation with the associated loss value for each token and the true and predicted labels. Predicted labels obtained with the CamemBERT-bio model on the RADIO corpus.

Input tokens	_de	_visible	_en	_un	_tasse	_du	_plateau	_spontané	_supérieur	_vis
count	988	14	278	123	11	396	3	4	90	121
mean	0.03	1.65	0.08	0.1	1.12	0.03	4.11	3.08	0.14	1.03
sum	25.84	23.06	21.28	12.59	12.33	12.33	12.33	12.32	12.31	12.31

Table 4
Mean, occurrence of tokens, and sum of losses for the ten input tokens representing the highest total losses in the validation set.

Labels	I-scope_neg	B-scope_neg	O	B-cue_neg	I-cue_neg
count	733	290	14,443	291	63
mean	0.32	0.09	0.01	0.0	0.0
sum	263.3	24.8	74.51	0.14	0.0

Table 5
Statistics of the loss values grouped by negation token labels showing the mean value of the losses for the tokens of the class, and their count in the validation dataset. The number in bold locates the larger mean loss value. Predicted labels obtained with the CamemBERT-bio model on the RADIO corpus.

to scope detection on all models. The model also sometimes fails to identify the beginning of the scope, though this represents a much lower mean loss value (0.09). The loss value for the other labels, in particular those related to cue detection are more than ten times lower, confirming that scope detection is a more difficult task than cue detection.

Next, we analyzed the confusion made by classifiers between tags. Figure 1 depicts a confusion matrix to facilitate visual interpretation. The off-diagonal values represent numbers of classification errors. We observe that the main source of error consists in confusing the inside of the scope with an O tag (11). Other types of error are actually very rare with this model.

Finally, we examined in detail some of the sentences that were incorrectly predicted by CamemBERT-bio on the RADIO corpus and represented noteworthy errors in our model. We found that a substantial amount of errors was due to tokenization errors. We show three representative examples below, with the real negation cue bolded, its real scope put in brackets and the predicted labels shown as top brackets. The tokens that are missed by the model are underlined. In the first sentence, the model correctly predicted the adjective “déplacées”, including the vowel "e" (feminine marker) as the beginning of the scope but ignored the vowel “s” (plural mark). During tokenization, the initial token “déplacées” was split into two sub-words, leaving an isolated “s”. Unlemmatized end-of-word vowels, such as “s”, are frequently associated with the majority entity O (token for “Other”) in the affirmative part of sentences, explaining the potential for confusion by the model. This type of error could be solved by lemmatizing the sentences before training to prevent the model from splitting the tokens. Example 2 illustrates another tokenization issue: the word “distendue” was split into three sub-words “dis”, “tendu” and “e”. The model labeled the first two tokens as the beginning of the scope and preferred the majority entity O to annotate the feminine marker "e". Finally, in the last examples, a partial negation marked by the indicator “sans”, which typically affects the first token, indicating the beginning of the scope. However, the model annotated the sequence “en contraste spontané” as the inside the scope, showing where the model can confuse the inside of the scope and the sequence of tokens representing the affirmative part of the sentence.

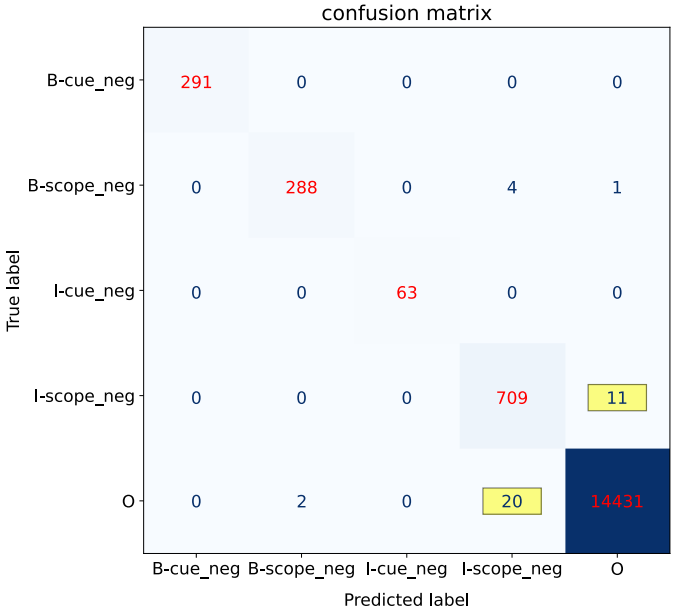

Figure 1: Confusion matrix for the labelling of the validation set of the RADIO corpus using the CamemBERT-bio model.

1. fracture **non** [déplacée s] des arcs postérieurs de k11 et k12 gauches.
2. vésicule biliaire **non** [dis tendu e], à paroi fine.
3. les anses digestives sont **sans** [particularité] en contraste spontané.

4.2. Results across different datasets

The excellent performance of the best models above for negation analysis on a corpus of thoracic CT scans is promising. However, as mentioned above, the context in which these results were achieved is more favorable than what would be expected from a practical implementation in the clinical routine. This section presents the results with validation across our three different datasets: RADIO, ESSAI and CAS.

Table 6 shows the performances obtained with the four scenarios, for the cue detection and the scope detection. Scenario 1 shows that the radiology-specific negation detection model obtained in section 3.1 above fails to detect some of the negation indices of ESSAI and CAS (F1-score of 92.9%). As a result, its performance for the detection of the scope is rather low (F1-score of 73.0%), thus confirming that the fine-tuning on the RADIO corpus does not provide the model with a strong capability for generalization. This result was expected because the ESSAI and CAS corpora contain many more negation indices than the RADIO corpora. Table 1 in the Appendix indeed shows that a large part of the negation cues of ESSAI and CAS does not appear in RADIO. In addition, several prediction errors we found to be due to improper scope identification. This in turn is due to the fact that the training (RADIO) and test (ESSAI+CAS) datasets were annotated by different annotators, with different annotation rules.

Under Scenario 2, all the models trained on the merging of ESSAI+CAS successfully identified the negation cues with very high precision (F1 $\geq$ 98.8%), a result that is again easily explained by analyzing the frequency of the negation indices in the training and test data (Table 1). The performances of the three models on the negation scope detection task is however much lower, with F1-scores that are between 16 and 20 lower than for cue detection. The best model for the

Scen	Models	TRAIN	TEST	Cue detection			Scope detection		
				P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
1	Neg-Radio-CamemBERT-bio-base	RADIO	ESSAI+CAS	92.0	93.7	92.9	66.7	80.6	73.0
2	CamemBERT-base	ESSAI+CAS	RADIO	98.6	99.3	98.9	84.7	80.1	82.3
	CamemBERT-bio-base			**98.8**	**99.4**	**99.0**	**85.5**	**80.3**	**82.8**
	DrBERT			98.3	99.2	98.8	80.4	77.6	79.0
3	CamemBERT-base	RADIO+ESSAI	CAS	95.0	94.6	94.8	83.3	**85.8**	84.5
	CamemBERT-bio-base			**95.8**	**94.7**	**95.3**	**84.7**	85.3	**85.0**
	DrBERT			94.5	93.2	93.8	77.1	78.8	78.0
4	CamemBERT-base	NLMFC	(10%) NLMFC	97.74 ± 0.47	$\mathbf{98.69 \pm 0.41}$	$\mathbf{98.21 \pm 0.27}$	90.40 ± 0.87	91.13 ± 0.69	90.47 ± 0.63
	CamemBERT-bio-base			$\mathbf{97.75 \pm 0.45}$	98.67 ± 0.20	98.20 ± 0.21	$\mathbf{90.48 \pm 0.74}$	$\mathbf{91.34 \pm 0.60}$	$\mathbf{90.90 \pm 0.58}$
	DrBERT			97.56 ± 0.45	98.40 ± 0.38	97.98 ± 0.27	$\underline{86.72 \pm 1.17}$	$\underline{88.36 \pm 1.06}$	$\underline{87.52 \pm 1.05}$

Table 6
Performance metrics (Precision, Recall, and F1-score) obtained with the four scenarios described in Section 3.3. The best models are shown in bold. Scenario 1 to 3 evaluate tokenization on an external dataset, so that we did not use cross-validation but present the results of the model fine-tuned on the complete training set. Thus the absence of indication of variance in these cases. For scenario 4, results are given as average±standard-deviation over the 10-fold cross-validation. Underlined numbers locate a statistically significant difference with the best model, using a t-test with a 5% significance level.

	Sentences	Sentences-lem
CamemBERT-bio-base	8,950.62	9,752.85
Neg-Radio-CamemBERT-bio-base	9,014.15	9,117.83
Ratio ↓	100.7%	93.5%

Table 7
Comparison of the mean perplexity computed over 11,037 lemmatized and non-lemmatized sentences from Corpus ESSAI+CAS for both models CamemBERT-bio-base and Neg-Radio-CamemBERT-bio-base. The lower the better. The third line illustrates the ratio of perplexities across rows (the lower the better ↓).

scenario is again CamemBERT-bio with F1-scores of 99.0% and 82.8% on the cue detection and scope detection tasks, respectively. This value of 82.8% for the scope may seem small, but keeping in mind that the training set (ESSAI+CAS) and the test set (RADIO) come from different contexts (generic biomedical texts vs reports of chest CT scans) and have been annotated in a quite heterogeneous way, this value should on the opposite be interpreted as a support to the robustness and generalizabilty of the models.

The purpose of Scenario 3 is the verification of the previous conclusions, by fine-tuning CamemBERT-base, CamemBERT-bio and DrBERT using the fusion of the RADIO and ESSAI corpora, then assessing its performance on the CAS corpus. Comparing the obtained results with Scenario 1 indicates that associating the ESSAI corpus with the RADIO training base in the training set has been highly beneficial. For CamemBERT-bio, this resulted in an increase of the F1-score by roughly 2.5 on cue detection and by 12 on scope detection (compared to scenario 1). Augmentation of the training data has enhanced the consistency and compatibility between the texts and their annotations. This underscores the significance of consistency between the data used for model training and the data used for model testing.

The results of Table 6 for Scenario 4, show that the three tested models exhibited very good performance on the cue detection task (around 98% for each model) with no significant differences between them. For scope detection, the performances remain lower than for cue detection (F1-scores decreased by 8 to 9) but, compared to all the previous scenarios, performance detection is now much larger. The best models (CamemBERT-bio and CamemBERT-base) achieved a F1-score above 90%, a performance well above all the models and scenarios of the table for negation scope detection.

4.3. Results of over-fitting audit

Table 7 provides the perplexity measures for the original CamemBERT-bio model and the fine-tuned version for negation detection (Neg-Radio-CamemBERT-bio). The third line of the table illustrates the ratio of perplexities between the original and the fine-tuned models. A ratio lower than 100% illustrates an improvement. This experiment

shows that our fine-tuned model has a very similar perplexity as the original model on raw sentences, and that the fine-tuning improved the perplexity by 6.5% when sentences are lemmatized.

5. Conclusion

In this article, we proposed the use of language models adapted to the medical field for the recognition of negation information in clinical texts written in French using the token classification approach. For this purpose, we used two public biomedical corpora annotated with negation information and a private corpus of radiological texts from the Hospices Civils of Lyon, which negations have been annotated by us. Five linguistic models have been fine-tuned for the detection of two specific classes (cue, scope) with the aim of identifying negation in thoracic CT scan reports, three of which are specifically tailored to the medical domain. The results of our approach indicates excellent performance for the detection of negation with the CamemBERT-bio model, with F1-scores of 99.4% and 94.5%, for the cue detection and scope detection tasks, respectively. These results compare favorably or improve on the previous approaches in the literature. For instance, NegBERT is a state-of-the-art proposal based on transformers for the detection of negation cues and scopes [17] in English corpora. On BioScope, a corpus of biomedical texts, NegBERT exhibited F1-scores between 90% and 96% (depending on the corpus subset) for cue detection and between 85% and 96% for the scope detection tasks. Our approach therefore exhibit better or comparable performance. Our examination of the main errors illustrates that using a tokenizer poorly suited to a specific domain can lead to errors in predicting out of vocabulary tokens. This highlights the significance of utilizing a tokenizer specifically tailored for radiology. Error analysis also reveals that the model may struggle to precisely delineate the scope when dealing with long sentences that involve total negation.

Whereas the first section of our study was restricted to reports of chest CT scans, we expanded it across more general medical and biomedical domains in the second part. We show that an increase in the size of the annotated dataset, the consistency between clinical texts from different medical specialties, and the annotation criteria, can have a fundamental impact on the performance of the negation detection system.

Three language models for the automatic processing of negations in french reports can be proposed from our study. First, the CamemBERT-bio-based model, fine-tuned on the RADIO corpus of chest radiology reports in French. Since this model is trained on private data, that may be sensitive in terms of privacy preservation, we cannot to date share the trained model freely. Likewise, the models trained according to Scenario 3 and 4 in Section 3.3 underwent training on a dataset that was in part composed of our private RADIO corpus. These models can therefore not be shared. However the CamemBERT-bio-based model trained on the merging of the ESSAI+CAS corpora achieved good performance on negation annotation, with F1-scores of 99.0 % for cue detection and 82.8% for scope detection. This model is public and can be freely downloaded on our Hugging Face page[6], where visitors can test its performance on input sentences.

Our results also evidence that that fine-tuning our model with negation did not lead to a catastrophic forgetting of the original capability of the model. Even better, when sentences were lemmatized, fine-tuning actually improved the quality of the model on its originak fill-mask taks. This shows that fine-tuning with a negation detection task improves the model ability to accurately capture the meaning of sentences in a biomedical corpus.

Finally, we note that a major interest of the approach used here is the correct handling of the polysemy of negation terms. For example, in the sentence: "Leur cytoplasme était fortement coloré par le PAS (Periodic-Acid-Schiff)", our approach correctly ignores the term "PAS", and does not considere it as a negation cue. The CamemBERT-bio-based model trained according to Scenario 4 in section 3.3 underwent training with over 4, 244 annotated negative sentences. It exhibited the highest performance of our study when multiple types of medical reports were used (radiology, clinical trials, biomedical cases, ...), even though the annotation rules were not consistent, with F1-scores of 98.2% and 90.9% for cue detection and scope detection, respectively. We believe that its performance and robustness with respect to the diversity of the annotation rules and the type of biomedical data will allow its utilization in a real-life clinical context.

Acknowledgments

The authors would like to thanks Anamaria Falaus, linguist with Nantes university, France, who helped us to better understand the subtleties and the diversity of negations cues in french. The valuable insights she provided helped us label our datasets correctly.

[6] https://huggingface.co/aistrosight/Neg-CamemBERT-bio

References

[1] S. Agarwal and H. Yu. Biomedical negation scope detection with conditional random fields. Journal of the American medical informatics association, 17(6):696–701, 2010.

[2] M. Ballesteros, V. Francisco, A. Díaz, J. Herrera, and P. Gervás. Inferring the scope of negation in biomedical documents. In Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLing), pages 363–375, 2012.

[3] A. Casey, E. Davidson, M. Poon, H. Dong, D. Duma, A. Grivas, C. Grover, V. Suárez-Paniagua, R. Tobin, W. Whiteley, H. Wu, and B. Alex. A systematic review of natural language processing applied to radiology reports. Medical Informatics and Decision Making, 21:179, 06 2021.

[4] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics, 34(5):301–310, 2001.

[5] S. F. Chen, D. Beeferman, and R. Rosenfeld. Evaluation metrics for language models, 1998.

[6] V. Cotik, V. Stricker, J. Vivaldi, and H. Rodriguez. Syntactic methods for negation detection in radiology reports in Spanish. In Proceedings of the Workshop on Biomedical Natural Language Processing, pages 156–165. Association for Computational Linguistics, 2016.

[7] N. P. Cruz Díaz. Negation and speculation detection in clinical and review texts. Procesamiento del Lenguaje Natural, 54:107–110, 2015.

[8] C. Dalloux, V. Claveau, and N. Grabar. Speculation and negation detection in french biomedical corpora. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), pages 223–232. INCOMA Ltd., 2019.

[9] C. Dalloux, V. Claveau, N. Grabar, L. E. S. Oliveira, C. M. Cabral Moro, Y. B. Gumiel, and D. R. Carvalho. Supervised learning for the detection of negation and of its scope in french and brazilian portuguese biomedical corpora. Natural Language Engineering, 27(2):181–201, 2021.

[10] L. Deléger and C. Grouin. Detecting negation of medical problems in french clinical notes. In Proceedings of the SIGHIT International Health Informatics symposium (IHI), pages 697–702. ACM, 2012.

[11] P. Delobelle, T. Winters, and B. Berendt. Robbert: a dutch roberta-based language model. arXiv preprint arXiv:2001.06286, 2020.

[12] F. Fancellu, A. Lopez, B. Webber, and H. He. Detecting negation scope is easy, except when it isn't. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL), pages 58–63. Association for Computational Linguistics, 2017.

[13] N. Grabar, C. Dalloux, and V. Claveau. CAS: corpus of clinical cases in french. Journal of Biomedical Semantics, 11(7), 2020.

[14] H. Harkema, J. N. Dowling, T. Thornblade, and W. W. Chapman. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of biomedical informatics, 42(5):839–851, 2009.

[15] S. M. Jiménez-Zafra, R. Morante, E. Blanco, M.-T. Martín-Valdivia, and L. A. U. López. Detecting negation cues and scopes in Spanish. In Proceedings of the Language Resources and Evaluation Conference (LREC), pages 6902–6911. European Language Resources Association, 2020.

[16] E. Jupin-Delevaux, A. Djahnine, F. Talbot, A. Richard, S. Gouttard, A. Mansuy, P. Douek, S. Si-Mohamed, and L. Boussel. BERT-based natural language processing analysis of French CT reports: Application to the measurement of the positivity rate for pulmonary embolism. Research in Diagnostic and Interventional Imaging, 6:100027, 2023.

[17] A. Khandelwal and S. Sawant. NegBERT: A transfer learning approach for negation detection and scope resolution. In Proceedings of the Language Resources and Evaluation Conference (LREC), pages 5739–5748. European Language Resources Association, 2020.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[19] I. Krsnik, G. Glavaš, M. Krsnik, D. Miletić, and I. Štajduhar. Automatic annotation of narrative radiology reports. Diagnostics, 10(4), 2020.

[20] Y. Labrak, A. Bazoge, R. Dufour, M. Rouvier, E. Morin, B. Daille, and P.-A. Gourraud. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 16207–16221. Association for Computational Linguistics, 2023.

[21] H. Li and W. Lu. Learning with structured representations for negation scope extraction. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 533–539. Association for Computational Linguistics, 2018.

[22] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot. CamemBERT: a tasty French language model. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 7203–7219. Association for Computational Linguistics, 2020.

[23] S. Mehrabi, A. Krishnan, S. Sohn, A. Roch, H. Schmidt, J. Kesterson, C. Beesley, P. Dexter, C. Schmidt, H. Liu, and M. Palakal. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. Journal of Biomedical Informatics, 54, 03 2015.

[24] R. Morante and W. Daelemans. A metalearning approach to processing the scope of negation. In S. Stevenson and X. Carreras, editors, Proceedings of the Conference on Computational Natural Language Learning (CoNLL), pages 21–29. Association for Computational Linguistics, 2009.

[25] R. Morante, A. Liekens, and W. Daelemans. Learning the scope of negation in biomedical texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (ENMLP), pages 715–724. Association for Computational Linguistics, 2008.

[26] A. Névéol, C. Grouin, J. Leixa, S. Rosset, and P. Zweigenbaum. The Quaero french medical corpus: a resource for medical entity recognition and normalization. In Proceedings of the Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing (BioTxtM), pages 24–30, 2014.

[27] A. Olthof, P. Van Ooijen, and L. Cornelissen. Deep learning-based natural language processing in radiology: The impact of report complexity, disease prevalence, dataset size, and algorithm type on model performance. Journal of Medical Systems, 45:91, 10 2021.

[28] O. S. Pabón, O. Montenegro, M. Torrente, A. R. González, M. Provencio, and E. Menasalvas. Negation and uncertainty detection in clinical texts written in spanish: a deep learning-based approach. PeerJ Computer Science, 8:e913, 2022.

[29] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. M. Summers, and Z. Lu. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. Proceedings of the AMIA Summits on Translational Science, 2018:188–196, 2018.

[30] Z. Qian, P. Li, Q. Zhu, G. Zhou, Z. Luo, and W. Luo. Speculation and negation scope detection via convolutional neural networks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 815–825. Association for

Computational Linguistics, 2016.
[31] S. Sohn, S. Wu, and C. Chute. Dependency parser-based negation detection in clinical narratives. Proceedings of the AMIA Summit on Translational Science, 2012:1–8, 03 2012.
[32] D. Sykes, A. Grivas, C. Grover, R. Tobin, C. Sudlow, W. Whiteley, A. Mcintosh, H. Whalley, and B. Alex. Comparison of rule-based and neural network models for negation detection in radiology reports. Natural Language Engineering, 27:1–22, 11 2020.
[33] G. Szarvas, V. Vincze, R. Farkas, and J. Csirik. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, pages 38–45. Association for Computational Linguistics, 2008.
[34] S. Taylor and S. Harabagiu. The role of a deep-learning method for negation detection in patient cohort identification from electroencephalography reports. Proceedings of the Annual AMIA Symposium, 2018:1018–1027, 12 2018.
[35] R. Touchent, L. Romary, and E. De La Clergerie. CamemBERT-bio: Un modèle de langue français savoureux et meilleur pour la santé. In Conférence en Recherche d'Information et Applications (CORIA), pages 323–334. ATALA, 2023.
[36] L. Tunstall, L. Von Werra, and T. Wolf. Natural language processing with transformers. O'Reilly Media, Inc., 2022.
[37] B. van Es, L. C. Reteig, S. C. Tan, M. Schraagen, M. M. Hemker, S. R. Arends, M. A. Rios, and S. Haitjema. Negation detection in Dutch clinical texts: an evaluation of rule-based and machine learning methods. Bioinformatics, 24(1):10, 2023.
[38] E. Velldal, L. Øvrelid, J. Read, and S. Oepen. Speculation and negation: Rules, rankers, and the role of syntax. Computational Linguistics, 38(2):369–410, June 2012.
[39] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. Atkinson, S. Amin, and H. Liu. A clinical text classification paradigm using weak supervision and deep representation. Medical Informatics and Decision Making, 19, 01 2019.
[40] K.-H. Weng, C.-F. Liu, and C.-J. Chen. Deep learning approach for negation and speculation detection for automated important finding flagging and extraction in radiology report: Internal validation and technique comparison study. JMIR Medical Informatics, 11:e46348, 2023.
[41] A. Yan, J. McAuley, X. Lu, J. Du, E. Y. Chang, A. Gentili, and C.-N. Hsu. RadBERT: Adapting transformer-based language models to radiology. Radiology: Artificial Intelligence, 4(4):e210258, 2022.
[42] J. Zech, M. Pain, J. Titano, M. Badgeley, J. Schefflein, A. Su, A. Costa, J. Bederson, J. Lehar, and E. Oermann. Natural language–based machine learning models for the annotation of clinical radiology reports. Radiology, 287:171093, 01 2018.
[43] Y. Zhai, S. Tong, X. Li, M. Cai, Q. Qu, Y. J. Lee, and Y. Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In Proceedings of the Conference on Parsimony and Learning, pages 202–227. PMLR, 2024.

6. Appendix

RADIO		ESSAI		CAS	
Cues negation	Occurences	Cues negation	Occurences	Cues negation	Occurences
(-)absence de/du/d'/des	244	absences de/du/d'	13	absence de/d'	51
absent	1	absentes	1	aucun(e)	10
aucun(e)	4	aucun....n'(e)	16	aucun(e)....n'	31
aucune....n'	1	aucun(e)	4	disparaissaient, disparition	36
à l'exception de/du	2	à l'exception d'/de/des	9	dépourvu, dépourvues	3
disparu	2	à la place	3	exclure, exclu	2
jamais	1	disparition	2	jamais	1
impossible	2	excepté(es)	29	impossible	3
non, -non	662	en dehors d'	1	n'...jamais	10
non plus	4	exclus, à l'exclusion	2	non plus	2
ni	143	hors	5	n'(ne)....aucun(e)	51
n'...aucune	5	hormis	1	négatif(s), négative(s)	33
n'...plus	4	impossibilité, impossible	22	ni	119
ne...jamais	1	incapacité	6	non	71
pas, ne (n')...pas	750	inaccessible(s)	3	n'...rien	3
sauf	2	n'...jamais	8	n'(ne)....plus	15
sans	1 493	non, non-	353	pas, n'(ne).....pas	361
		ni	91	sans	279
		n'(ne)...plus	5		
		négatif,négative	11		
		nulle	1		
		n'(ne)...aucun	6		
		pas, n'(ne)...pas	276		
		sans	188		

Table 1
Frequency of the main negation indicators within the RADIO, ESSAI and CAS corpora.

	RADIO	ESSAI	CAS
Sentence numbers	10,798	7,247	3,790
Numbers of tokens	146,471	138,964	72,493
Numbers of negative sentences	2,321	981	837
Numbers of negative indices (total)	3,291	1,056	1,011
Numbers of negative indices (unique)	19	29	23

Table 2
Statistical summary for the RADIO, ESSAI and ESSAI corpora.