

Domain-Specific Long Text Classification from Sparse Relevant Information

Célia D’Cruz^{a,*}, Jean-Marc Bereder^{b,**}, Frédéric Precioso^{a,***} and Michel Riveill^{a,****}

^aUniversité Côte d’Azur, CNRS, Inria, I3S, France

^bCHU de Nice, France

Abstract. Large Language Models have undoubtedly revolutionized the Natural Language Processing field, the current trend being to promote one-model-for-all tasks (sentiment analysis, translation, etc.). However, the statistical mechanisms at work in the larger language models struggle to exploit the relevant information when it is very sparse, when it is a weak signal. This is the case, for example, for the classification of long domain-specific documents, when the relevance relies on a single relevant word or on very few relevant words from technical jargon. In the medical domain, it is essential to determine whether a given report contains critical information about a patient’s condition. This critical information is often based on one or few specific isolated terms. In this paper, we propose a hierarchical model which exploits a short list of potential target terms to retrieve candidate sentences and represent them into the contextualized embedding of the target term(s) they contain. A pooling of the term(s) embedding(s) entails the document representation to be classified. We evaluate our model on one public medical document benchmark in English and on one private French medical dataset. We show that our narrower hierarchical model is better than larger language models for retrieving relevant long documents in a domain-specific context.

1 Motivations and related works

With the impact of Large Language Models (LLMs) over a wider community than the Natural Language Processing (NLP) one, even beyond the boundaries of the scientific community, there is a tendency to explore the full capabilities of these new models in particular for every NLP task. It seems that "bigger is always better" and "one-model-for-all-tasks" (e.g. the foundation models) are the new trendy mottos in NLP.

Domain-specific and technical jargon issues Despite some remarkable successes of LLMs, when investigating the potential of these approaches for specific tasks, a very recent work [25] has shown that smaller language models are more appropriate in a domain-specific context, especially when high accuracy is expected because we are considering a high-stakes domain, such as law in their paper or medicine in our work. In their work, the authors consider more specifically several algorithms based on BERT backbone [6], RoBERTa [15], DistilBERT [23]. The authors explain the impact of considering a technical jargon, called "*Legalese language*" in Law

domain, where words which are similar in general English cannot be considered as similar by the model because they are two distinct categories (of crime) in legal texts, as for instance theft and fraud [25]. Such configurations are obvious challenges for very large language models pre-trained on a huge amount of data in general English.

We face the exact same problems with the technical jargon present in the medical reports. Patient medical reports contain a significant amount of information on diverse health characteristics in structured text such as key-value pairs (e.g. lab results) or in unstructured free text (e.g. medical history). However, specific information contained in free text can be hard to retrieve. Reports can often be several pages long, and similar terms in general English can appear in different places of the same report confusing the language model on the final relevance of the report. Examples can be found in the literature: although anterior and ventral are similar in meaning, anterior is used to indicate spatial relationships of body parts to each other throughout the body, but ventral is most often used to indicate a relationship to the anterior abdominal wall; the use of upper rather than superior has evolved from common usage rather than following a strict anatomical designation. For instance, upper lip and upper eyelid are proper clinical designations rather than superior lip or superior eyelid, but superior pole of the kidney is the proper terminology for the "upper" or "top" of the kidney.

We also face the dual phenomenon that the wording in free text may widely differ from one report to another, both in the linguistic structure and in the medical terms. Different medical terms, as well as Latin form, or medical abbreviations (sometimes even codes), can be used to describe similar patient’s condition [18, 17]: asthenia or weakness, arthralgia or joint pain, epistaxis or nosebleed, etc.

We do not even mention typos that can transform *ileum* (the last section of the small intestine between the jejunum and the colon) into *ilium* (a part of the pelvic bone), or *peroneal* (pertaining to the lateral aspect of the leg) into *perineal* (pertaining to the pelvic floor).

All of these issues on a rigorous use of a precise medical terminology and their consequences on health care are well documented and taken seriously in training programs in medicine ¹.

This context explains why our first attempts to investigate recent advances in information retrieval based on language models have not been conclusive [11, 21]. Still, the methods proposed in these papers are based on variants of BERT backbone (the aforementioned ones, or ALBERT [13], ERNIE [34]), as for the previously cited papers.

* Corresponding Author. Email: celia.dacruz@inria.fr

** Corresponding Author. Email: bereder.jm@gmail.com

*** Corresponding Author. Email: frederic.precioso@inria.fr

**** Corresponding Author. Email: michel.riveill@inria.fr

¹ European Medicine Agency, Medical terms simplifier, https://www.ema.europa.eu/en/documents/other/ema-medical-terms-simplifier_en.pdf

Long text classification issues It took only a few years for Transformer-based models [30] to become state-of-the-art for text encoding, drastically improving text classification performances. However, when focusing on the aforementioned encoders most of them are limited to 512 tokens as the maximum input sequence length, limiting the amount of text that can be taken into account. These text embedding methods have been mostly used for short text classification, while medical reports can be several pages long.

For classifying long texts, several strategies have been proposed. The most basic strategies involve indiscriminately truncating long documents, thus completely discarding large text parts that can contain relevant information. For instance, taking into account only the beginning, or the end, or a random part of a medical report.

Other models involve selecting key sentences from documents such as CogLTX [7] and Bert+TextRank [20].

Other approaches adapt existing pre-trained transformer models by using hierarchical strategies [32, 8, 19, 4], dividing the whole text into small parts that are iteratively fed to the transformer model.

Another strategy involves the use of long-range Transformers [26, 27] that employ various strategies to reduce quadratic self-attention complexity, thus enabling them to have a significantly higher limit in the number of tokens that can be processed (e.g. Longformer [2], Reformer [12], BigBird [33]). However, most of these developments have been primarily made for English.

Finally, the most recent approaches that have been proposed are based on extracting embedding layers from decoder architectures such as Llama 2 [28].

However, when the objective is to focus on a specific critical detail, as opposed to topic classification, relevant information represents only a small portion of all the information present in the reports. The relevant information may appear a few times in a report, or even only once. Additionally, the available training corpus size can be quite limited. In this context, most of the existing models above do not demonstrate higher performances for our peculiar classification while requesting significantly bigger computation resources.

Privacy and computation resources concerns In [24], Schick and Schütze show that small language models (ALBERT, RoBERTa) can be few-shot learners as larger models have shown. The authors mention that they achieve similar performances as larger LLMs with “*language models that are much “greener” in that their parameter count is several orders of magnitude smaller.*” In [10], Hsieh et al. emphasizes how large LLMs are “*challenging to deploy in real world applications due to their sheer size*” and that their “*computational requirements are far beyond affordable for most product teams, especially for applications that require low latency performance.*”

This concern is particularly relevant in our case since working on medical personal data prevents from outsourcing computations using cloud solutions, and almost no health center hosts computation resources is able to run very large language models on a regular basis. Designing a lighter model is therefore one of our objectives, and to do this, relying on one of the algorithms in the BERT family is a reasonable choice.

In this paper, we present our deep learning approach designed to classify medical reports, documents ranging from one to several pages in length, focusing on critical but sparse specific information, swamped with noisy and confusing content.

We propose three main contributions:

- A filtering phase to reduce the representativeness gap between relevant and irrelevant content based on a short list of target terms (up

to 30), from the semantic field of the information to be retrieved. For reproducibility sake, we provide alternatives to generate this short list from publically available external resources. We show that our deep model is not much impacted by the size of an automatically generated list even if it contains extra useless terms.

- A hierarchical deep model based, first, on the BERT embedding of the potentially relevant target terms contained in the filtered sentences, then on the attention weight average of these term embeddings to result in the document embedding that is finally classified.
- An extensive evaluation and ablation study conducted against all the state-of-the-art approaches (other hierarchical deep models, non-hierarchical models, long-range transformers), showing also the robustness of our model with respect to the quality of the short list of target terms used during the filtering phase, and demonstrating the overall performance of our solution.

In this paper we detail both the private French corpus and the public English benchmark used for evaluation (Sec. 2), describe our approach (Sec. 3), and test the effectiveness of our approach against reference approaches on the two benchmarks (Sec. 4).

2 Datasets

2.1 Public English dataset for smoking status

We use the publicly available English dataset “2006 -Smoking” [29], where the classification task is to identify patient smoking status from medical discharge records. This dataset was released as part of the National NLP Clinical Challenges (n2c2, previously i2b2) ².

The dataset consists of 502 de-identified discharge summaries annotated by pulmonologists to classify patient reports into five categories: *Past Smoker* (patient is a former smoker who has not smoked for at least one year), *Current Smoker* (patient is currently a smoker or was a smoker within the past year), *Smoker* (a Current or a Past Smoker but the report does not provide enough information to classify as either), *Non-Smoker* (patient has never been a smoker), *Unknown* (the report does not mention anything about smoking). There is an average of 874 words per report, with a maximum of 3292.

Some reports do not give any indication on the patient smoking status. In other reports, while this information is provided, it is mentioned once or at very few places, therefore representing only a tiny fraction of a report. Moreover, the wording indicating the patient status varies from one report to another. For instance, both sentences “She is a former smoker” and “Nicotine abuse, quit in the 80s” indicates that the patient is a past smoker.

The provided dataset contains 104 reports as testing data (11 *Past Smoker*, 11 *Current Smoker*, 3 *Smoker*, 16 *Non-Smoker*, 63 *Unknown*) and 398 reports as training data (36 *Past Smoker*, 35 *Current Smoker*, 9 *Smoker*, 66 *Non-Smoker*, 252 *Unknown*).

2.2 French dataset from colorectal cancer patients

The French dataset³ consists of 198 de-identified French medical reports from colon cancer patients obtained from Nice University Hospital. All identifying information - including patients’ and doctors’ names, birthdays, ID numbers, locations - has been removed. The

² Website for the National NLP Clinical Challenges (n2c2) <https://n2c2.dbm.i.hms.harvard.edu/>

³ This dataset is private and has been created under an agreement between Nice University Hospital and Université Côte d’Azur.

anonymity was verified through the use of regex rules and each report was carefully read multiple times by medical experts. The medical records cover the period from the beginning of 2017 to the end of 2021. They range from one page to a few pages long (572 words on average per report, maximum 2115) with a pre-determined sequence of sections. However, the text of each section is freely written. One of the main types of reports is follow-up letters with sections describing the patient's co-morbidities, cancer history, previous treatments, biological parameters, etc. Other report types are pathological reports, surgical reports, imaging results, etc.

Those reports may contain information about the *laterality* of the primary colon tumor (classified as either "left", "right", or "no information"), whether a *surgery* for the cancer has already been performed (classified as either "yes" or "no / no information"), and whether a *chemotherapy* has already been initiated (classified as either "yes" or "no / no information"). Some reports do not indicate any of this information, while other reports indicate several instances of this information in one or more sections.

Our dataset has been annotated by a surgical oncologist and a Deep Learning (DL) expert in two ways: the exact locations of the relevant information in each report, as well as the overall classification of the information at the report level. As an example, for the *laterality* of the primary colon tumor, if a report indicates "The patient has a tumor in the left colon and a metastasis in the right lobe of the liver.", only the word "left" is annotated as indicating the *laterality* of the primary colon tumor, since the word "right" does not indicate the *laterality* of the primary tumor, it refers instead to a metastasis. The whole report is classified as "left" for the *laterality*.

Out of 198 reports, the document label distribution for *laterality* is 67 (right), 81 (left), and 50 (no). The distribution for *surgery* is 44 (no) and 154 (yes) and for *chemotherapy* it is 145 (no) and 53 (yes).

Although the private French medical dataset cannot be shared, we still present the results as it is interesting in the sense that the list of relevant keywords was manually defined by medical experts, and that the classification task *laterality* is particularly challenging since "left" and "right" are target terms but are also everywhere in reports.

3 Approach

3.1 Overview

For contexts where the relevant information appears quite sparsely in the documents, models that indiscriminately take into account the whole text will likely struggle to focus on the few relevant parts to accurately classify the documents. The idea of our approach is to only keep the potentially relevant parts of a document while discarding most of the remaining probably irrelevant parts, therefore enabling our transformer-based deep learning model to focus only on a fraction of a document.

The relevant parts of a document are retrieved by building a list of target terms potentially relevant to the classification task, and then by extracting all the sentences containing one or more of those target terms. Figure 1 illustrates our approach.

3.2 Relevant target term list

We compile a list of target terms relevant to each classification task. It should be noted that those target terms can be single words or composed of several words.

3.2.1 English dataset relevant terms

The list of relevant terms aims at identifying the patient smoking status in the English dataset. We test two different approaches. In the first approach, the relevant terms are automatically extracted from the website relatedwords.io⁴ that gathers, from a single source target term, a list of related terms sorted by relevance. In our case, since we want to classify the smoking status of patients (Current Smoker, Past Smoker, Smoker, Non-Smoker, Unknown), we choose the source target term "smoker". In the second approach, we do not use any external resource; we use the X-Class [31] keyword expansion method with the source target term "smoker" to retrieve relevant terms related to "smoker" within the dataset from the training documents themselves.

We remove English stop words from the aforementioned list of relevant target terms using the Python library `stopwords`⁵ that removes terms such as "not", "non", "very". Therefore, the final list contains the source target term "smoker" and its related terms from the relatedwords.io website minus the stopwords.

This final target term list is ranked by the relevance of the terms to the words "smoker" in both relatedwords.io website and X-Class methods. However, we can vary the list size by removing the terms ranked as least relevant in order to study the sensitivity of our model to the target term list, as we did in our experiments.

3.2.2 French dataset relevant terms

The list of relevant target terms for each classification task – *laterality* (of the primary colon tumor), *surgery* (performed), and *chemotherapy* (initiated) – to identify relevant information from the French dataset has been provided by the medical experts from Nice University Hospital, who considered that list as more relevant than using a French version of relatedwords.io⁶. The three lists provided by the experts range from approximately 10 to 20 terms. Those lists are much smaller than the ones extracted from relatedwords.io because those three lists have been carefully tailored to each classification task, which are quite specific. Those lists provided by the experts are not ranked by relevance. We simulate variable list sizes by manually adding terms that are of little relevance to the classification task (namely "augmented vocabulary"), or manually deleting some relevant terms from those lists (namely "reduced vocabulary").

3.3 Filtering

Once the list of relevant target terms has been established, we search in the documents for all instances of those target terms.

Case A: if a document does not contain any target term, it is categorized as "no / no information" regarding *laterality*, *surgery*, or *chemotherapy*, or "unknown" regarding the patient smoking status. No further action is taken.

Case B: if target terms are present, we extract all sentences from the documents that include these target terms. These sentences provide a context. These target terms may be relevant, because they actually provide explicit information on the label of the document for one of the target classes (e.g. the *laterality*, the occurrence of a *surgery*, or the initiation of *chemotherapy*), but they may also be irrelevant ("past shoulder surgery after road accident" does not indicate a *surgery* for cancer, "the use of a *chemotherapy* treatment will later be discussed"

⁴ <https://relatedwords.io>

⁵ <https://pypi.org/project/stopwordsiso/>

⁶ For instance : <https://www.rimessolidess.com/motscles.aspx?m=chirurgie>

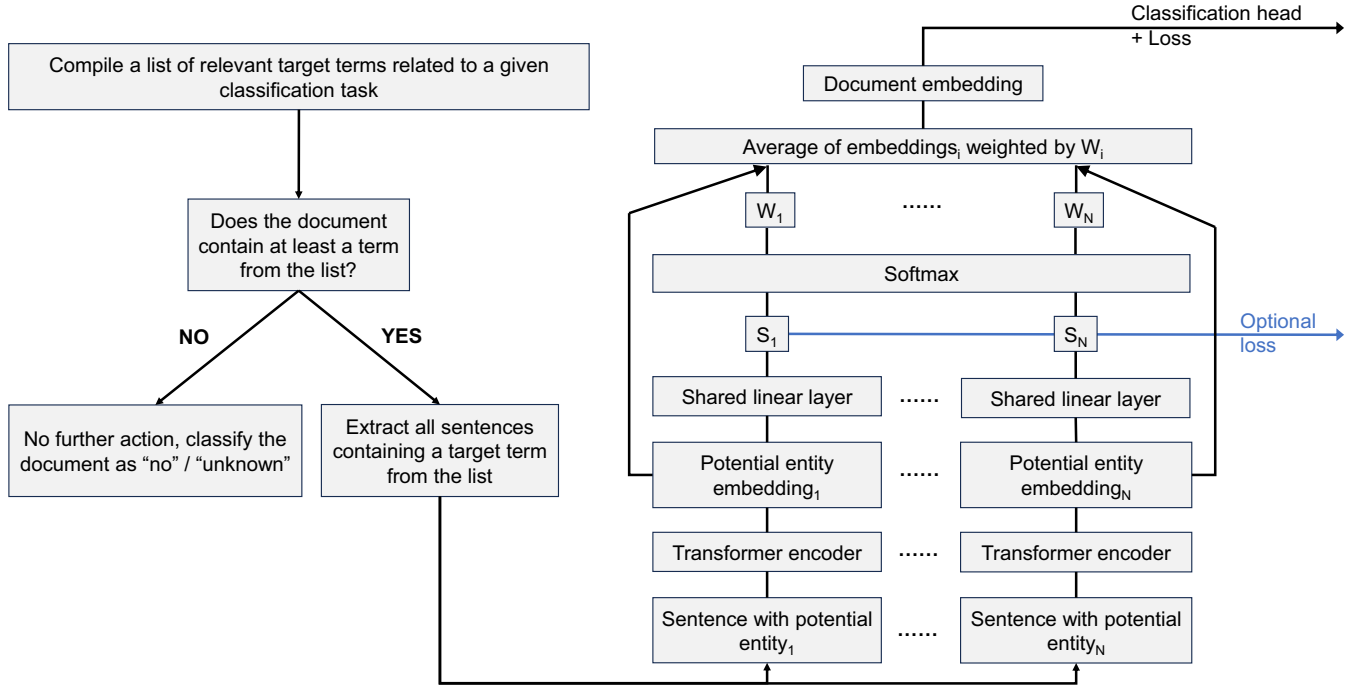


Figure 1. Architecture of our approach

does not indicate that a chemotherapy has already been initiated). The target term "smoker" can be used in sentences such as "she is a former smoker", "she is not a smoker", "current smoker", which does not lead to the same smoking status label. From now on, we call *entities* the target terms in the context of their surrounding sentence.

3.4 Model architecture

3.4.1 Main architecture

The filtering done in the previous step retained the sentences containing the relevant vocabulary. We obtain a contextualized embedding of each target term by passing the sentence to which it belongs to a transformer encoder and selecting specifically the embedding associated to the target term.

Some entities may actually be irrelevant to the classification task, while others are highly relevant. The model calculates an attention weight W_i for each target term to give more weights to relevant entities. This is done by passing each target term embedding to a shared linear layer that outputs a score S_i and normalizing all the scores with a softmax layer.

The document embedding is the product sum of each entity embedding and their attention weight. This vector is then used in the classification head that includes a linear layer followed by a softmax layer with cross-entropy loss function.

3.4.2 Model extension

While a first filtering phase has been performed to select potentially relevant sentences based on the presence of target terms, it might still be challenging in certain cases to obtain very good classification results. For instance, the laterality classification task is challenging because the target terms "right" and "left" may very often appear within a report but be relevant to the primary tumor laterality only

in a few sentences. Therefore, a second filtering phase can be optionally performed by submitting the sentences in the training set initially selected from the first filtering phase to the experts for further manual annotation to determine whether the terms within their sentence are indeed relevant or not. However, to reduce the need of interactions with the experts, an active learning strategy can be employed to annotate only a small fraction of the documents during this second filtering phase.

Our initial model architecture can be extended by adding a sigmoid function after the S_i scores followed by an optional binary cross entropy loss to leverage the annotation of the second optional filtering phase at training time. Table 7 shows the improved performance on our three document classification tasks in our French dataset that has been further annotated for the second optional filtering phase.

4 Results

4.1 Baseline model and DL contenders

We benchmark our approach on both our French and English datasets against Machine Learning (ML) and DL architectures designed for long text classification.

- **ML model.** Logistic regression (LR) where we removed beforehand stopwords and punctuations from the reports, we lemmatized and changed the text to lowercase, and applied TF-IDF.
- **Sentence selection DL model.** Bert+TextRank [20].
- **Flat DL model.** "SAN"-like model [9] that performs classification on the document embedding obtained by simply performing a weighted average of all subword embeddings.
- **Hierarchical DL models.** 1) "HiSAN"-like model [9] with a Transformer encoder to replace their non-contextual embeddings and we split the texts into blocks of 512 tokens. 2) "Transformer over CLS" with max pooling [3].

	Laterality		Surgery		Chemotherapy	
	Accuracy	Balanced Accuracy	Accuracy	Balanced Accuracy	Accuracy	Balanced Accuracy
LR	0.556	0.542	0.843	0.753	<u>0.874</u>	<u>0.872</u>
HiSAN	0.526	0.490	0.847	0.761	0.866	0.841
Transformer over CLS	0.561	0.543	0.866	0.737	0.851	0.820
SAN	0.579	0.544	0.860	<u>0.866</u>	0.839	0.834
Bert+TextRank	<u>0.689</u>	<u>0.662</u>	<u>0.878</u>	0.787	0.865	0.862
Ours	0.782	0.741	0.940	0.892	0.925	0.875

Table 1. Model comparisons on individual tasks on our French dataset

- **Long-range transformers.** 1) Clinical-LongFormer using global attention only on the first token [14]. 2) Clinical-BigBird using sparse attention [14]. 3) Llama 2 7B [28] using the embedding of the last token. If a text exceeds 4096 tokens, only the beginning of the text up to 4096 tokens is used as input to the models.

4.2 Experimental setup

Except for ML and long-range transformer approaches, we use the Bio+Discharge Summary BERT [1] pretrained model from Hugging Face library⁷ as the Transformer encoder for the English dataset, and the CamemBERT-base pretrained model [16] from Hugging Face library⁸ as the Transformer encoder for our French dataset. We use PySBD [22] to identify sentence boundaries.

A 5-fold cross-validation procedure was used on the French dataset, splitting the data into three fifths for training, one fifth for validation, and one fifth for test. For the English dataset, the test set was already defined in the National NLP Clinical Challenges (n2c2), the 5-fold cross-validation led to four fifths for training and one fifth for validation. For both datasets, the validation sets were used only for setting the Early Stopping, and the results were averaged over the 5 folds. The hyperparameters were not set using the cross-validation procedure, but set as commonly reported in the literature. We only had to reduce the batch size from 32 to 4 owing to restricted computational resources.

For each classification task and for each architecture, we train a separate model. Our evaluation metrics are the accuracy⁹ and balanced accuracy¹⁰. In all DL experiments, we use the same optimizer (AdamW), effective batch size (4), learning rate (2e-5), early stop patience (5) and label smoothing (0.1).

4.3 Experimental results

4.3.1 Comparison with diverse types of approach

We compare our approach with different types of models as illustrated in Table 1 on the French dataset and Table 2 on the English dataset. Our approach shows substantially better results than ML, flat, or other sentence selection models. Likewise, our narrower hierarchical model, which uses a fraction of the text by only selecting potentially relevant sentences, demonstrates better results than other

	Smoking status	
	Accuracy	Balanced Accuracy
LR	0.596	0.215
Bert+TextRank	0.604	0.253
SAN	0.817	0.551
Transformer over CLS	0.848	0.624
HiSAN	0.823	0.577
Llama 2	0.829	0.560
LongFormer	0.813	0.599
BigBird	0.875	0.652
Ours (N = 30) relatedWords.io	0.900	<u>0.665</u>
Ours (N = 20) X-Class	<u>0.887</u>	0.668

Table 2. Model comparisons on the English dataset. "Ours" corresponds to our approach with the first N most relevant target terms extracted from the relatedwords.io website or X-Class method

hierarchical models where the whole text is indiscriminately fed iteratively to a Bert-like model. Long-range transformers allow the entire documents to be processed all at once, but they are computationally expensive and our lighter approach shows better performances.

4.3.2 Robustness of the vocabulary list

Both the French and English datasets show that our approach is not very sensitive to noise in the target term list, namely the vocabulary list. The inclusion of additional unnecessary terms or the omission of non-critical terms does not have a major impact on the results.

Table 3 shows, on the English dataset, the impact of the variation of the number N of most relevant target terms generated either by the relatedwords.io website or the X-Class keyword expansion method.

With the relatedwords.io approach on the English dataset, the first N = 10 target terms does not include some critical terms such as "smoking", hence resulting in poor results. The first N = 20 target terms does include "smoking" as well as "nonsmoker", therefore substantially improving the results. The first N = 30 target terms includes additionally relevant terms and little noise, giving the best results.

⁷ https://huggingface.co/emilyalsentzer/Bio_Discharge_Summary_BERT

⁸ <https://huggingface.co/camembert-base>

⁹ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

¹⁰ https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html

	relatedwords.io		X-Class	
	Accuracy	Balanced Accuracy	Accuracy	Balanced Accuracy
Ours (N = 10)	0.788	0.489	0.887	0.658
Ours (N = 20)	0.894	0.655	0.887	0.668
Ours (N = 30)	0.900	0.665	0.879	0.659
Ours (N = 40)	0.887	0.665	0.867	0.639
Ours (N = 50)	0.887	0.658	0.871	0.640

Table 3. Variation of the number N of most relevant target terms on the English dataset

	Laterality	
	Accuracy	Balanced Accuracy
Ours	0.782	0.741
Augmented vocabulary	0.752	0.701
Reduced vocabulary	<u>0.764</u>	<u>0.723</u>

Table 4. Vocabulary list variations for the laterality classification task

	Surgery	
	Accuracy	Balanced Accuracy
Ours	<u>0.940</u>	0.892
Augmented vocabulary	0.927	0.874
Reduced vocabulary	0.942	<u>0.887</u>

Table 5. Vocabulary list variations for the surgery classification task

	Chemotherapy	
	Accuracy	Balanced Accuracy
Ours	0.925	0.875
Reduced vocabulary	<u>0.921</u>	<u>0.871</u>
Further Reduced vocabulary	0.902	0.841

Table 6. Vocabulary list variations for the chemotherapy classification task

Increasing N only slightly decreases the performances as noise increases with less relevant terms.

With the X-Class target term expansion method on the English dataset, the first N = 10 terms already includes critical terms such as "smoking", "tobacco", "cigarette", therefore already showing very good results. Using N = 20 shows similar results, and using additionally less relevant terms only slightly decreases the performance.

Tables 4, 5 and 6 show the impact of the variation of the vocabulary lists on the French dataset. "Reduced vocabulary" lists only include the most relevant target terms that are present in all the relevant documents, "Ours" lists include additionally relevant terms, while "augmented vocabulary" lists include additional terms that are of little relevance. The supplementary material [5] details the vocabulary lists used in the French dataset. For instance, in the case of the laterality classification task, our "reduced vocabulary" list only contains the "left" and "right" target terms as they are the most present in the reports. "Ours" list adds the target terms related to anatomical parts of the colon that indicates the laterality ("caecum", "sigmoid", etc.). Our "augmented vocabulary" list contains a few additional terms that are related to the colon but are not relevant in the reports we have, such as "transverse" (middle part of the colon, neither left or right).

Furthermore, our approach is not very sensitive to the method used to find the target terms: interviewing experts from the domain, using the X-Class keyword expansion method, or using the relatedwords.io website yield strong results as long as the vocabulary lists contain enough critically relevant terms.

4.3.3 Impact of the sparsity of the relevant information

Models that indiscriminately consider the whole text may struggle to classify documents when the relevant information appears sparsely and the rest of the text offers very little hints. The impact the degree of sparsity is best illustrated by Table 1 on the French dataset. For the laterality task, the structure or type of reports (e.g. follow-up letters, imaging reports, etc.) provide very limited clues on the laterality of the colon tumor, and the information of the colon tumor laterality appears at most only in a few sentences in reports that can be a few pages long, making it challenging for other models to classify the documents, whereas our approach shows a large improvement. However, for the chemotherapy and surgery tasks, certain types of reports offer some clues. For instance, pathological reports are often related to a surgery and some types of follow-up letters are usually related to a chemotherapy. In this case, our model still shows better performance, but the improvement over other models is less substantial.

4.3.4 Optional second loss function

The colon tumor laterality classification task is particularly challenging since "right" and "left" terms appear everywhere in the reports and are often not relevant (e.g. "right" or "left" in a text may refer to the colon tumor laterality which is relevant, but also to the side of a lung metastasis which is irrelevant). To cope with this issue, we have thus evaluated the impact of a further refinement process involving a second targeted annotation of the sentences selected through the first

	Laterality		Surgery		Chemotherapy	
	Accuracy	Balanced Accuracy	Accuracy	Balanced Accuracy	Accuracy	Balanced Accuracy
Ours	0.782	0.741	0.940	0.892	0.925	0.875
Ours + 2nd loss	0.904	0.891	0.979	0.970	0.940	0.901

Table 7. Document classification results of our model with the optional second loss function

relevant term list filtering, in the training set only. We have used this refined annotation with an additional loss function to help the model training on more precisely “relevant” / “non relevant” terms and sentences from the text. Since this require extra human operations, we only evaluated this process on the French dataset.

Training the model with both losses (the initial classification one, and the refinement filtering one where all filtered sentences have been further annotated in the training set), our approach shows, in Table 7, significant improvements for all three classification tasks compared to using only the classification loss function, especially for the *laterality* classification task as expected but still significantly also for *surgery* and *chemotherapy*.

An interesting future work would be designing an active learning strategy to annotate only a small subset of filtered sentences, requesting little work from domain experts for a significant improvement.

5 Conclusion

In this paper, we present an approach to identify if long documents are useful for experts based on relevant information which appears sparsely in those documents, as it is the case for instance in medical reports. We tackle this task considering it as a binary classification task: the document analyzed contains or not the relevant information.

Traditional ML and DL models designed for long text classification struggle in this context. Our model is based on hierarchical representations combining token-level embeddings with a sentence-level representation into a document embedding allowing to classify precisely the document as relevant or irrelevant. This specific structure combining different embedding levels of the text outperforms long range transformers as well as hierarchical deep models that have been proposed in the literature for long document classification.

Text embedding in our model is based on BERT variants, nowadays so called small language models, and outperforms more recent LLM-based embeddings, hence confirming recent recurrent results in similar contexts that larger LLMs are not always better.

We even propose a human (expert) in the loop mechanism for contexts where the vocabulary is confusing based on too general and broad target terms, improving the performances even more. Our approach outperforms SOTA on both a very specific French dataset and a more classic public English benchmark, and can easily be extended to other classification tasks and datasets.

In future work, we plan to investigate the extra load of refining annotation of selected sentences through an active learning strategy.

6 Ethical considerations

Our work holds the potential to prevent healthcare professionals from unintentionally overlooking critical details in patients’ files, therefore potentially enhancing clinical decision-making. However, our model is not flawless, and the mistakes it makes could conversely negatively impact clinical decisions. Hence, our work should complement, not

replace, expert judgment. This research should be seen as a step towards facilitated information extraction while emphasizing the irreplaceability of human expertise in medical contexts.

Moreover, our private French dataset is de-identified to ensure patient confidentiality and complies with General Data Protection Regulation and the strict rules of processing medical data in France. The raw data are not processed out of the hospital, but concerns over data privacy and potential misuse could still be raised. We are aware of those concerns and all this work is conducted under a strict privacy and confidentiality procedure of the hospital.

References

- [1] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott. Publicly available clinical BERT embeddings. In A. Rumshisky, K. Roberts, S. Bethard, and T. Naumann, editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.
- [2] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020. URL <https://arxiv.org/abs/2004.05150>.
- [3] X. Dai, I. Chalkidis, S. Darkner, and D. Elliott. Revisiting transformer-based models for long document classification. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7212–7230. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDING-S-EMNLP.534.
- [4] X. Dai, I. Chalkidis, S. Darkner, and D. Elliott. Revisiting Transformer-based Models for Long Document Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7212–7230, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.534. URL <https://aclanthology.org/2022.findings-emnlp.534>.
- [5] C. D’Cruz, J.-M. Bereder, F. Precioso, and M. Riveill. Domain-specific long text classification from sparse relevant information. *arXiv preprint arXiv:2408.13253*, 2024. Full version of this paper.
- [6] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL <http://arxiv.org/abs/1810.04805>.
- [7] M. Ding, C. Zhou, H. Yang, and J. Tang. Coglitx: Applying BERT to long texts. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/96671501524948bc3937b4b30d0e57b9-Abstract.html>.
- [8] S. Gao, M. T. Young, J. X. Qiu, H.-J. Yoon, J. B. Christian, P. A. Fearn, G. D. Tourassi, and A. Ramanathan. Hierarchical attention networks for information extraction from cancer pathology reports. *Journal of the American Medical Informatics Association*, 25(3):321–330, Mar. 2018. ISSN 1067-5027, 1527-974X. doi: 10.1093/jamia/ocx131. URL <https://academic.oup.com/jamia/article/25/3/321/4636780>.
- [9] S. Gao, J. X. Qiu, M. Alawad, J. D. Hinkle, N. Schaefferkoetter, H.-J. Yoon, B. Christian, P. A. Fearn, L. Penberthy, X.-C. Wu, L. Coyle, G. Tourassi, and A. Ramanathan. Classifying cancer pathology reports with hierarchical self-attention networks. *Artificial Intelligence in Medicine*, 101:101726, Nov. 2019. ISSN 09333657. doi: 10.1016/j.artmed.2019.101726. URL <https://linkinghub.elsevier.com/retrieve/pii/S0933365719303562>.
- [10] C.-Y. Hsieh, C.-L. Li, C.-K. YEH, H. Nakhost, Y. Fujii, A. J. Ratner, R. Krishna, C.-Y. Lee, and T. Pfister. Distilling step-by-step! outper-

- forming larger language models with less training data and smaller model sizes. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.
- [11] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550.
 - [12] N. Kitaev, L. Kaiser, and A. Levskaya. Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
 - [13] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
 - [14] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, and Y. Luo. A comparative study of pretrained language models for long clinical text. *J. Am. Medical Informatics Assoc.*, 30(2):340–347, 2023.
 - [15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL <http://arxiv.org/abs/1907.11692>.
 - [16] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, E. De La Clergerie, D. Seddah, and B. Sagot. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.645. URL <https://www.aclweb.org/anthology/2020.acl-main.645>.
 - [17] K. McCaffery, B. Nickel, R. Moynihan, J. Hersch, A. Teixeira-Pinto, L. Irwig, and A. Barratt. How different terminology for ductal carcinoma in situ impacts women’s concern and treatment preferences: a randomised comparison within a national community survey. *BMJ Open*, 5(11), 2015. ISSN 2044-6055. doi: 10.1136/bmjopen-2015-008094. URL <https://bmjopen.bmj.com/content/5/11/e008094>.
 - [18] B. Nickel, A. Barratt, T. Copp, R. Moynihan, and K. McCaffery. Words do matter: a systematic review on how different terminology for the same condition influences management preferences. *BMJ Open*, 7(7), 2017. ISSN 2044-6055. doi: 10.1136/bmjopen-2016-014129. URL <https://bmjopen.bmj.com/content/7/7/e014129>.
 - [19] R. Pappagari, P. Żelasko, J. Villalba, Y. Carmiel, and N. Dehak. Hierarchical transformers for long document classification. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 838–844. IEEE, 2019.
 - [20] H. H. Park, Y. Vyas, and K. Shah. Efficient classification of long documents using transformers. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 702–709. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-SHORT.79. URL <https://doi.org/10.18653/v1/2022.acl-short.79>.
 - [21] Y. Qu, Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, and H. Wang. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, 2021.
 - [22] N. Sadvilkar and M. Neumann. Pysbd: Pragmatic sentence boundary disambiguation. *CoRR*, abs/2010.09657, 2020. URL <https://arxiv.org/abs/2010.09657>.
 - [23] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
 - [24] T. Schick and H. Schütze. It’s not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, 2021.
 - [25] R. Sheik, K. Siva Sundara, and S. J. Nirmala. Neural data augmentation for legal overruling task: Small deep learning models vs. large language models. *Neural Processing Letters*, 56(2):121, 2024.
 - [26] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler. Long range arena : A benchmark for efficient transformers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
 - [27] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler. Efficient Transformers: A Survey. *ACM Computing Surveys*, 55:1–28, July 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3530811. URL <https://dl.acm.org/doi/10.1145/3530811>.
 - [28] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/ARXIV.2307.09288. URL <https://doi.org/10.48550/arXiv.2307.09288>.
 - [29] O. Uzuner, I. Goldstein, Y. Luo, and I. Kohane. Identifying Patient Smoking Status from Medical Discharge Records. *Journal of the American Medical Informatics Association*, 15(1):14–24, Jan. 2008. ISSN 1067-5027, 1527-974X. doi: 10.1197/jamia.M2408.
 - [30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
 - [31] Z. Wang, D. Mekala, and J. Shang. X-class: Text classification with extremely weak supervision. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3043–3053. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAAACL-MAIN.242. URL <https://doi.org/10.18653/v1/2021.naacl-main.242>.
 - [32] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1174. URL <http://aclweb.org/anthology/N16-1174>.
 - [33] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontañón, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed. Big bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
 - [34] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. ERNIE: enhanced language representation with informative entities. In A. Korhonen, D. R. Traum, and L. Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics, 2019.