Diagnostic surveillance of high-grade gliomas: towards automated change detection using radiology report classification

Tommaso Di Noto[1 [0000−0002−5161−055X]], Chirine Atat[1 [0000−0003−2097−821X]], Eduardo Gamito Teiga[1 [0000−0001−6677−4808]], Monika Hegi[2,3,4 [0000−0003−0855−6495]], Andreas Hottinger[5 [0000−0001−7098−9414]], Meritxell Bach Cuadra[1,6 [0000−0003−2730−4285]], Patric Hagmann[1 [0000−0002−2854−6561]], and Jonas Richiardi[1 [0000−0002−6975−5634]]

[1] Department of Radiology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[2] Neuroscience Research Center, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[3] Neurosurgery, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[4] Swiss Cancer Center Léman (SCCL), Lausanne, Switzerland
[5] Department of Clinical Neurosciences; Department of Oncology, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[6] Medical Image Analysis Laboratory, Center for Biomedical Imaging, Lausanne, Switzerland

Abstract. Natural Language Processing (NLP) on electronic health records (EHRs) can be used to monitor the evolution of pathologies over time to facilitate diagnosis and improve decision-making. In this study, we designed an NLP pipeline to classify Magnetic Resonance Imaging (MRI) radiology reports of patients with high-grade gliomas. Specifically, we aimed to distinguish reports indicating changes in tumors between one examination and the follow-up examination (treatment response/tumor progression versus stability). A total of 164 patients with 361 associated reports were retrieved from routine imaging, and reports were labeled by one radiologist. First, we assessed which embedding is more suitable when working with limited data, in French, from a specific domain. To do so, we compared a classic embedding techniques, TF-IDF, to a neural embedding technique, Doc2Vec, after hyperparameter optimization for both. A random forest classifier was used to classify the reports into stable (unchanged tumor) or unstable (changed tumor). Second, we applied the post-hoc LIME explainability tool to understand the decisions taken by the model. Overall, classification results obtained in repeated 5-fold cross-validation with TF-IDF reached around 89% AUC and were significantly better than those achieved with Doc2Vec (Wilcoxon signed-rank test, $P = 0.009$). The explainability toolkit run on TF-IDF revealed some interesting patterns: first, words indicating change such as *progression* were rightfully frequent for reports classified as unstable; similarly, words indicating no change such as *not* were frequent for reports classified as stable. Lastly, the toolkit discovered misleading words such as *T2*

which are clearly not directly relevant for the task. All the code used for this study is made available.

Keywords: Natural Language Processing (NLP) · Term Frequency - Inverse Document Frequency (TF-IDF) · Doc2Vec · diagnostic surveillance · LIME Model Explainability

1 Introduction

In the last decade, Machine Learning (ML) has reshaped research in radiology. ML models yield state-of-the-art results for numerous medical imaging tasks such as segmentation, anomaly detection, registration, and disease classification [1]. In addition to images, ML models have also been increasingly applied to radiology reports and more generally to data coming from Radiology Information Systems (RIS) [2]. However, even though radiology reports contain valuable, high-level insights from trained physicians, they also come with some associated drawbacks; in particular, most reports are stored as unstructured, free-text documents. Consequently, they exhibit a strong degree of ambiguity, uncertainty and lack of conciseness [3].

Natural Language Processing (NLP) is a branch of ML that helps computers understand, interpret, and manipulate human language [4]. In the case of radiology reports, NLP has the goal of extracting clinically relevant information from unstructured texts. As recently illustrated in one extensive review [5], one frequent application of NLP for radiology reports is diagnostic surveillance. Its objective is to monitor the evolution of a pathology in order to extrapolate useful knowledge and improve decision-making. In line with this trend, our work focuses on oncology patients with high-grade gliomas that are scanned longitudinally for frequent follow-up.

According to [5], the majority (86%) of studies published up until 2019 focused on medical reports written in English, while only 1% of the reviewed studies utilized French reports. This language gap is understandable given that a substantial portion of NLP tools was developed using English texts. Nonetheless, in medical NLP, researchers need to adapt their models to the language of the radiology reports. This entails custom precautions and expedients to take since languages are often syntactically and/or semantically different from English. In this work, we investigate NLP methods for radiology reports written in French.

In addition, [5] concluded that although a growing number of Deep Learning (DL) NLP methods has been applied in recent years, "conventional ML approaches are still prevalent". To assess which technique is more suitable for our dataset, we compare two traditional embedding strategies, namely Term Frequency–Inverse Document Frequency (TF-IDF) [6] and Doc2Vec [7].

The task that we address is binary document classification. Specifically, we aim to identify the main conclusion of the medical reports deciding among the following groups: tumor *stability* vs. tumor *instability*. Details about these classes

are provided in section 2.2. The potential applications of our report classifier are twofold: first, it could help referring physicians to focus the attention on the main conclusion of the report, thus accelerating subsequent decisions. Second, the predicted classes could be used as weak labels for a downstream machine learning task (e.g. automated cohort creation). In addition, most clinically relevant images in RIS are associated with a radiology report, and thus offer potential access to several hundred thousands of weakly labelled images in medium to large hospitals.

In this work we also conduct an interpretability analysis of the model's decisions [8,9], based on the post-hoc interpretation technique LIME [10]. Its main objective is to identify the most important words that influenced the final prediction, by creating a surrogate linear model that performs local input perturbation (details in section 2.4).

In summary, this study presents a classifier for French radiology reports in the context of diagnostic surveillance, while comparing two embedding techniques and providing a visual interpretation of the model's decisions.

1.1 Related Works

Here, we present the works most similar to ours. In [11], the authors compared several embedding techniques and five different classifiers for detecting the radiologist's intent in oncologic evaluations. Similarly, [12] investigated a DL model to identify oncologic outcomes from radiology reports. The authors in [13] utilized a combination of ML and rule-based approaches to highlight important changes and identify significant observations that characterize radiology reports. [14] devised a model that extracts radiological measurements and the corresponding core descriptors (e.g. temporality, anatomical entity, ...) from Magnetic Resonance (MR), Computed Tomography (CT) and mammography reports. The work of [15] describes an NLP pipeline that identifies patients with (pre)cancer of the cervix and anus from histopathologic reports. Last, [16] detected thromboembolic diseases and incidental findings from angiography and venography reports.

Among all these works, only [16] used French reports, while the others worked with English documents. Moreover, only [12] addressed the issue of model explainability which we believe is paramount for the ML community, especially in the medical domain.

2 Materials and Methods

2.1 Dataset

We retrospectively included 164 subjects that underwent longitudinal MR glioma follow-up in the university hospital of Lausanne (CHUV) between 2005 and 2019. 71% of the patients in the cohort had Glioblastoma Multiforme (GBM), while

the remaining 29% had either an oligoastrocytoma or an oligodendroglioma. At every session, a series of MR scans were performed including structural, perfusion and functional imaging. For the sake of this study, we only focused on the native T1-weighted (T1w) scan, the T2-weighted (T2w) scan and the T1w-gad (post gadolinium injection, a contrast agent). For 25 patients, we collected images and reports across multiple sessions (on average, 9 sessions per subject). For the remaining 139 patients, we only retrieved images and reports from 1 random session. This latter sampling strategy was adopted to increase the chance of having cases of tumor progression and tumor response, since multiple sessions of the same subject mostly showed tumor stability and thus led to a very imbalanced data set. Overall, we ended up with a dataset of 361 radiology reports to use for the NLP pipeline. Every report was written (dictated) during routine clinical practice by a junior radiologist after exploring all sequences of interest. Then, a senior radiologist reviewed each case amending the final report when necessary. The extracted reports have varying length ranging from 114 to 533 words (average 255, standard deviation 68). The MR acquisition parameters for the cohort are provided in Table 1. The protocol of this study was approved by the regional ethics committee; written informed consent was waived.

Table 1: MR acquisition parameters of scans used for the study population.

# sessions ≡ # reports	**Vendor**	**Scanner**	**Field Strength [T]**
174	Siemens Healthcare	Skyra	3.0
73	Philips	Intera	3.0
46	Siemens Healthcare	Prisma	3.0
32	Siemens Healthcare	Symphony	1.5
21	Siemens Healthcare	TrioTim	3.0
10	Siemens Healthcare	Aera	1.5
5	Siemens Healthcare	Verio	3.0

2.2 Report Tagging

In order to build a supervised document classifier, one radiologist (4 years of experience in neuroimaging) tagged the reports with labels of interest. For each report, the annotator was instructed to perform two separate tasks: first, she had to assign 3 classes to the reports; one class that indicated the global conclusion of the report, one class to indicate the evolution of the enhanced part of the lesion (T1w conclusion) and the last one to indicate the evolution of the lesion on T2-weighted sequences (T2w conclusion). For each of these three groups, the annotator could choose between the following labels:

– **Stable**: assigned when the tumor did not change significantly with respect to the previous comparative exam.

– **Progression**: assigned when the tumor worsened with respect to the previous comparative exam. This class included cases where the enhanced part of the tumor increased in size or when the T2 signal anomalies surrounding the tumor increased in extension.
– **Response**: assigned when the tumor responded positively to the treatment (either chemotherapy or radiotherapy).
– **Unknown**: used when the annotator was not able to assign any of the three classes above.

The second task of the annotator was to highlight the most recent comparative date in the reports. Since the reports are not structured, this helped linking the current report being tagged with the most meaningful previous one. For simplicity, in this work we only focused on the global conclusion of the reports, and not on the T1 and T2 conclusions. Also, we removed all cases that were tagged as **unknown** (21 reports) and we merged **progression** and **response** into one unique class which we denote as **unstable**. By doing this, we narrowed the task to a binary classification problem where the model tries to distinguish between **stable** and **unstable** reports. After these modifications, we ended up with 191 **stable** reports and 149 **unstable** reports.

To facilitate the annotation process, we utilized the open-source software Dataturks[1]. This provided a graphic interface to the annotator which allowed her to tag, skip, highlight, and review the reports in a user-friendly way. Moreover, it automatically generated machine-readable labels once the annotation process was over. One exemplary report is illustrated in Figure 1, together with the corresponding annotations.

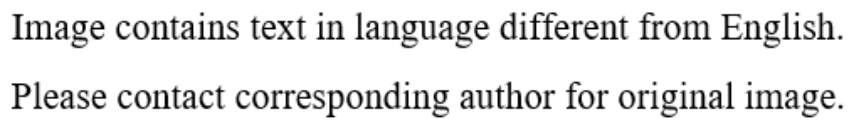

Fig. 1: Dataturks annotation interface. The annotator can select the classes in the left box and highlight the text of interest. Sensitive information has been blacked out for privacy.

2.3 Text Preprocessing & Embedding

Several preprocessing steps were carried out to reduce the vocabulary size. First, we removed all proper nouns such as physicians' and patients' names. This was

[1] OpenSource Data Annotation tool - `http://github.com/DataTurks/DataTurks`

performed using a pre-trained French Part-Of-Speech tagger from the Spacy library (version 3.0.6) [19]. Second, all the words in the reports were converted to lowercase. This operation is typical when there are no words that indicate a specific meaning when expressed with capital letters. Third, we removed punctuation and the most common French stop words, namely ['de', 'la', 'en', 'et', 'du', 'd', 'le', 'l', 'un', 'une', 'les', 'des', 'ces', 'à', 'au', 'aux']. Among these, we ensured to keep the French negation 'pas' (*not*) since it is very frequent in the reports, and reverses the meaning of the sentence. Fourth, all reports were tokenized using the *wordpunct* class of the Natural Language Toolkit framework (version 3.6.1) [20]. As last step, since all the reports contain the three sections '*indications*', '*description*' and '*conclusion*', we removed all content before the '*indication*' section, which is either useless (e.g. department phone number) or sensitive (e.g. patient identifier).

A key step in any NLP pipeline is text embedding. This corresponds to the conversion of tokenized text into numerical vectors. Historically, many embedding techniques have been proposed in literature. In this work, we compare two of the most widespread, namely TF-IDF [6] and Doc2Vec [7]. While the former is a standard term-weighting embedding scheme (traditional ML) that preserves the length of the tokenized documents, the latter is a DL-based technique that creates dense vectors which encode word order and context. TF-IDF was performed at the word level with the sklearn package (version 0.24.1) [21], whereas Doc2Vec was performed using the gensim library (version 4.0.1) [22].

2.4 Experiments

All experiments were run in a 5-fold, nested, stratified cross validation (CV). The internal CV was used to tune the hyperparameters of the pipeline with a custom Grid Search algorithm. Instead, the external CV was used to compute results on hold-out test samples. For TF-IDF, two hyperparameters were tuned: first, the types of retained N-grams were searched in the range [3,5]. Second, the percentage of vocabulary size to use was varied between 100% (all words are used) and 90% (the 10% rarest words are removed). The other parameters were fixed: the minimum document frequency was set to 2 and the maximum document frequency was set to 0.9 (indicating 90% of the documents). For Doc2Vec, the algorithm type (PV-DM or PV-DBOW) and the vector dimensionality [10, 30, 50] were tuned with the validation set. The context window was set to 5 words. Five "noise" negative words were drawn. Words with a total frequency lower than 2 were ignored. The model was trained for 100 epochs. Since stop words are not necessarily useless for Doc2Vec, we also tried to run the Doc2Vec pipeline preserving them.

The stratification of the CV guaranteed that both training and test sets contained approximately the same percentage of reports indicating tumor **stability** and tumor **instability**. To avoid overoptimistic predictions, we also ensured that the reports from multiple sessions of the same subject were not present some in the train set and some in the test set. Furthermore, to reduce the bias introduced by the random choice of patients at each CV split, the whole nested CV was

repeated 10 times, each time performing the splitting anew, and results were averaged.

For all experiments, we adopted the Random Forest algorithm [23] to classify the embedded documents, using once again the sklearn package. As hyperparameters, we set a fixed number of 501 trees and we tuned the maximum retained features in the internal CV, choosing between 0.8 (only 80% of the features are used) and 1.0 (all features are used).

To compare the two pipelines (Doc2Vec vs. TF-IDF embedding), we computed all standard classification metrics, namely accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1-score. Moreover, we also plotted the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves. The reports indicating tumor stability were considered as negative samples, whereas those indicating a change in the tumor were considered as positive samples. The classification metrics and the curves were averaged across the 10 runs. To statistically compare the classification results, a Wilcoxon signed-rank test was performed [24]. For simplicity, the test only accounted for the area under the ROC curve (AUC) across the 10 runs. A significance threshold level $\alpha = 0.05$ was set for comparing P values.

The explainability analysis was performed with the LIME toolkit on the TF-IDF pipeline only since it resulted in higher performances (see Table 2). We set the best hyperparameters obtained across the random runs and we ran LIME over all test reports. For each report, the toolkit performs a post-hoc interpretation following a two-step approach: first, it randomly generates neighborhood data in the vicinity of the example being explained; then, it "learns locally weighted linear models on this neighborhood data to explain each of the classes in an interpretable way". The user can choose how many features (words) are shown in the explanation. For this work, we set a maximum of 6 features per document. These weighted features represent the linear model which approximates the behaviour of the random forest classifier in the vicinity of the explained test example.

All the Python 3.6 code developed for this study is available on github[2].

3 Results

3.1 Classification performances

The nested CV with the Doc2Vec embedding took 50 minutes per run, while the one with TF-IDF took 2 hours. The most frequent hyperparameters chosen in the internal CV for Doc2Vec across the 10 random runs were a vector size of 10 and the PV-DV version of the algorithm. Instead, for TF-IDF, n-grams in the range (1,3) were the most frequent, and the optimal percentage of vocabulary size was 90%. For the Random Forest classifier, the configuration with 80% of the features was most frequent.

[2] https://github.com/connectomicslab/Glioma_NLP

We report in Table 2 the classification results of the two pipelines (TF-IDF vs. Doc2Vec), averaged over the 10 runs. Similarly, figures 2 and 3 illustrate the average ROC and PR curves. When comparing the two pipelines across the 10 random runs with the Wilcoxon signed-rank test, the AUC values of TF-IDF were significantly higher than those of Doc2Vec ($P = 0.009$). Last, classification results of the Doc2Vec pipeline run preserving the stop words led to higher results (average AUC = .85± .03). However, these were still significantly lower than the TF-IDF pipeline.

Table 2: Classification results across the 10 random runs. Values are presented as mean ± standard deviation. Bold values indicate the highest performances. Acc = accuracy; Sens = sensitivity; Spec = specificity; PPV = positive predictive value; NPV = negative predictive value; F1 = F1-score; AUC = area under the ROC curve; AUPR = area under the PR curve.

Embedding	Acc %	Sens %	Spec %	PPV %	NPV %	F1 %	AUC	AUPR
TF-IDF	**88**±1	91±1	**75**±0	**95**±0	**60**±2	**93**±0	**.89**±.01	**.97**±.00
Doc2Vec	86±2	**94**±3	38±10	89±1	57±10	92±1	.83±.05	.96±.01

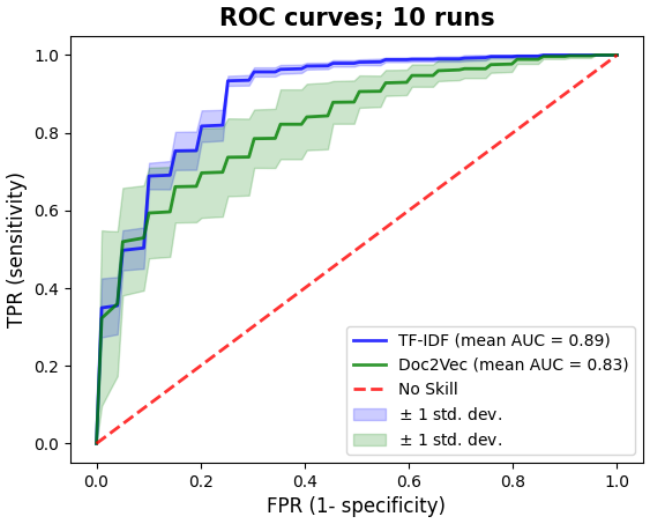

Fig. 2: Receiver operating characteristic (ROC) curves of the two pipelines (TF-IDF vs. Doc2Vec) averaged across the 10 runs.

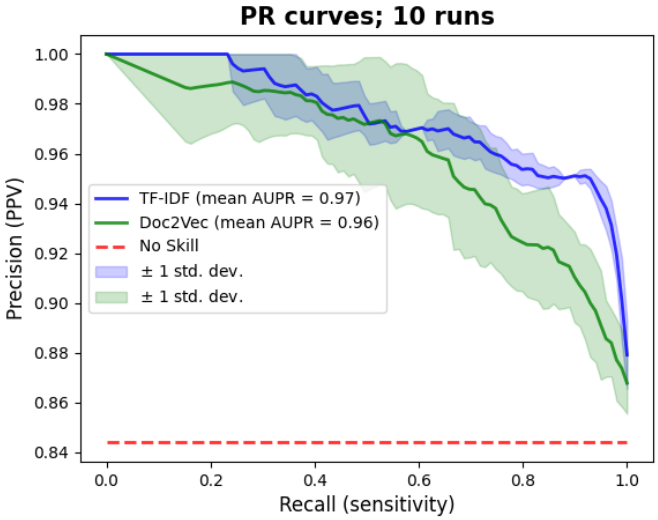

Fig. 3: Precision-Recall (PR) curves of the two pipelines (TF-IDF vs. Doc2Vec) averaged across the 10 runs.

3.2 Error Analysis & Model Interpretation

To further understand the decisions taken by the random forest algorithm, we applied the LIME post-hoc interpretability toolkit. Specifically, we investigated both the explanations created for the correctly classified reports and for the false positive and false negative reports. Table 3 shows the most frequent words used by the linear classifier created by LIME. We notice that most of the words intuitively make sense for the True Positive and True Negative samples. For instance, words like '*progression*', '*augmentation*' and '*diminution*' that all indicate some sort of change are recurrent for predicting TP samples and outweigh the corresponding words indicating tumor stability such as '*sans*' (*without*) or '*récidive*' (*recurrence*). A similar trend can be observed for TN samples where words like '*pas*' (*not*), '*stabilité*' (*stability*) and '*inchangé*' (*unchanged*) outweigh words indicating instability like '*apparition*' (*appearance*). However, the error analysis also highlighted some recurrent mistakes, such as the importance given to the words '*t2*' and '*axial*' in the FN samples or '*2007*' in the FP which ultimately deteriorate the predictions. To have a qualitative idea of the output of the LIME toolkit, we show in Figures 4 and 5 one TP and one FN example, respectively.

4 Discussion

In this work, we explored the potential of NLP for the task of diagnostic surveillance in patients with high-grade gliomas. As pointed out in [5], and subsequently shown in other works [25,26], traditional ML embedding techniques can

Table 3: Six most frequent features (words) used by the linear model generated by LIME to predict the class of the reports, sorted in descending order. For instance, the word '*progression*' is the most frequent word indicating instability used by the linear classifier for the TP test documents, whereas '*pas*' (French negation) is the most frequent word indicating stability used for the TN test documents. TP = True Positive (i.e. reports indicating tumor instability and predicted as such); TN = True Negative; FP = False Positive; FN = False Negative.

	Stable	**Unstable**
TP	sans	progression
	récidive	augmentation
	pas	oedème
	signe	plus
	anomalie	diminution
	ou	spectroscopie
TN	pas	apparition
	récidive	augmentation
	sans	axial
	stabilité	spectroscopie
	transverse	plus
	inchangé	postérieure

	Stable	**Unstable**
FP	sans	progression
	depuis	axial
	appareil	diminution
	réalisé	plus
	inchangé	oedème
	pondération	2007
FN	récidive	apparition
	pas	spectroscopie
	sans	augmentation
	transverse	diminution
	t2	axial
	stabilité	dans

lead to comparable results with respect to DL techniques when properly tuned. Moreover, they are still frequent when the dataset size is limited such as in medical imaging applications. Our work confirms this trend since, given the same classifier, the TF-IDF pipeline statistically outperformed the Doc2Vec one. The explainability analysis highlighted interesting trends. For the correctly classified reports, it confirmed that the model is focusing on relevant words. When investigating reports indicating instability, most of the recurrent terms indeed indicate a status of change such as '*diminution*', '*progression*' or '*plus*' (*more*). Similarly, the recurrent words for the reports indicating tumor stability reflect a status of no-change (e.g. '*pas*' (French negation)). Regarding the errors of the model, the LIME toolkit also uncovered some misleading words which obfuscate the final predictions. For instance, the words '*appareil*' (*MR scanner*), '*t2*', '*axial*' or '*transverse*' are recurrent in the explanations of FP and FN even though they are related to the acquisition process rather the status of the tumor.

The following limitations must be acknowledged. First, the annotations were performed by one single radiologist which is not the optimal scenario for ambiguous NLP tasks. Second, the dataset size is still limited with respect to similar studies [11,12,14].

In future works we are planning to enlarge the dataset and add a second annotator to assess inter-rate variability (and ideally intra-rater variability as well). Also, we would like to investigate which part of the report is the most important with respect to the final prediction. For instance, we would like to evaluate classification performances when using only *description* and *conclusion* of the reports, or even just the *conclusion*. In addition, we are planning to ex-

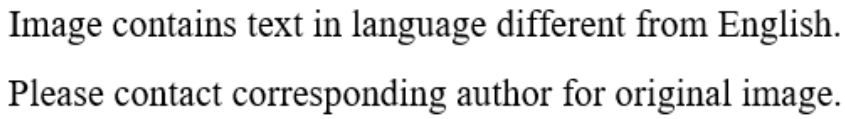

Fig. 4: LIME toolkit explanations for a TP report. Words such as '*diminution*' and '*augmentation*' correctly outweigh words indicating stability like '*pas*' (French negation) or '*sans*' (*without*). Sensitive information has been blacked out for privacy.

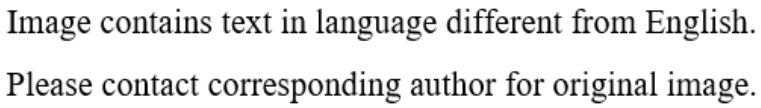

Fig. 5: LIME toolkit explanations for a FN report. Words such as '*sans*' and '*transverse*' incorrectly outweigh the key word indicating instability in this report which is '*apparition*' (*appearance*). Sensitive information has been blacked out for privacy.

periment different classifiers, or French pre-trained embedding models developed with larger corpora. Next, we will investigate what happens when shifting from a binary problem (*stable* vs. *unstable*) to a more granular task. Last, we will leverage the information extracted by the explainability toolkit to further preprocess the documents, for instance removing terms related to the acquisition protocol.

In conclusion, this work presented an NLP pipeline for the classification of radiology reports for patients with high-grade gliomas. The top-performing model (TF-IDF + Random Forest) attained satisfactory performances (AUC = .89) that lays a good foundation for generating weak labels, and the post-hoc explainability toolkit that we used holds promise for the development of a robust and transparent ML analysis.

References

1. Shen, Dinggang, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis." Annual review of biomedical engineering 19 (2017): 221-248.
2. Lakhani, Paras, et al. "Machine learning in radiology: applications beyond image interpretation." Journal of the American College of Radiology 15.2 (2018): 350-359.
3. Schwartz, Lawrence H., et al. "Improving communication of diagnostic radiology findings through structured reporting." Radiology 260.1 (2011): 174-181.
4. Chowdhury, Gobinda G. "Natural language processing." Annual review of information science and technology 37.1 (2003): 51-89.
5. Casey, Arlene, et al. "A Systematic Review of Natural Language Processing Applied to Radiology Reports." arXiv preprint arXiv:2102.09553 (2021).
6. Sammut, Claude, and Geoffrey I. Webb, eds. "Encyclopedia of machine learning." Springer Science & Business Media, 2011.
7. Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International conference on machine learning. PMLR, 2014.
8. Lipton, Zachary C. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue 16.3 (2018): 31-57.
9. Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).
10. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.
11. Chen, Po-Hao, et al. "Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports." Journal of digital imaging 31.2 (2018): 178-184.
12. Kehl, Kenneth L., et al. "Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports." JAMA oncology 5.10 (2019): 1421-1429.
13. Hassanpour, Saeed, Graham Bay, and Curtis P. Langlotz. "Characterization of change and significance for clinical findings in radiology reports through natural language processing." Journal of digital imaging 30.3 (2017): 314-322.
14. Bozkurt, Selen, et al. "Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm." Journal of digital imaging 32.4 (2019): 544-553.
15. Oliveira, Carlos R., et al. "Natural Language Processing for Surveillance of Cervical and Anal Cancer and Precancer: Algorithm Development and Split-Validation Study." JMIR medical informatics 8.11 (2020): e20826.
16. Pham, Anne-Dominique, et al. "Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings." BMC bioinformatics 15.1 (2014): 1-10.
17. Carletta, Jean. "Assessing agreement on classification tasks: the kappa statistic." arXiv preprint cmp-lg/9602004 (1996).
18. Gwet, Kilem L. "Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters." Advanced Analytics, LLC, 2014.
19. Honnibal, Matthew, et al. and Montani, "spaCy: Industrial-strength Natural Language Processing in Python", Zenodo, 2020, https://doi.org/10.5281/zenodo.1212303

20. Bird, Steven, Ewan Klein, and Edward Loper. "Natural language processing with Python: analyzing text with the natural language toolkit." O'Reilly Media, Inc.", 2009.
21. Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.
22. Rehurek, Radim, and Petr Sojka. "Gensim–python framework for vector space modelling." NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic 3.2 (2011).
23. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
24. Wilcoxon, Frank. "Individual comparisons by ranking methods." Breakthroughs in statistics. Springer, New York, NY, 1992. 196-202.
25. Dessi, Danilo, et al. "TF-IDF vs word embeddings for morbidity identification in clinical notes: An initial study." arXiv preprint arXiv:2105.09632 (2021).
26. Marcińczuk, Michał, et al. "Text document clustering: Wordnet vs. TF-IDF vs. word embeddings." Proceedings of the 11th Global Wordnet Conference. 2021.