Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)
© 2022 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI220605

Classification of Oncology Treatment Responses from French Radiology Reports with Supervised Machine Learning

Jean-Philippe GOLDMAN[a,d,1], Luc MOTTIN[b,c], Jamil ZAGHIR[a,d], Daniel KESZTHELYI[a,d], Belinda LOKAJ[a], Hugues TURBÉ[a,d], Julien GOBEIL[b], Patrick RUCH[b,c], Julien EHRSAM[a] and Christian LOVIS[a,d]
[a] *Division of Medical Information Sciences, Geneva University Hospital, Switzerland*
[b]*HES-SO/HEG Genève, Information Sciences, Geneva, Switzerland*
[c]*SIB Text Mining, Swiss Institute of Bioinformatics, Switzerland*
[d] *Department of Radiology and Medical Informatics, University of Geneva, Switzerland*

Abstract. The present study shows first attempts to automatically classify oncology treatment responses on the basis of the textual conclusion sections of radiology reports according to the RECIST classification. After a robust and extended manual annotation of 543 conclusion sections (5-to-50-word long), and after the training of several machine learning techniques (from traditional machine learning to deep learning), the best results show an accuracy score of 0.90 for a two-class classification (non-progressive vs. progressive disease) and of 0.82 for a four-class classification (complete response, partial response, stable disease, progressive disease) both with Logistic Regression approach. Some innovative solutions are further suggested to improve these scores in the future.

Keywords. oncology, treatment response, RECIST, automatic classification, supervised machine learning, natural language processing

1. Introduction

The main objective of the SPO project (Swiss Personalized Oncology) funded by the SPHN (Swiss Personalized Health Network) is to develop nation-wide infrastructure for personalized oncology and to maximize benefit for the patients. A part of this consists in developing a continuous, high quality clinical and molecular data collection throughout Switzerland as well as providing tools for the best personalized care for patients.

Standardizing radiology reports for the evaluation of response to treatment across institutes implies to establish a pipeline mining unstructured texts in the electronic health record (EHR) and to extract knowledge such as diagnosis, tumor type, primary tumor site, response to treatment, and morphologic description (texture, dimension, location).

This paper describes the efforts to automatically classify the response to treatment from the conclusion section of radiology reports in French from the Geneva University Hospitals (HUG) and the Lausanne University Hospital (CHUV). The two main goals are to annotate existing EHRs with the RECIST classification (Response Evaluation

[1] Corresponding Author, Jean-Philippe GOLDMAN; E-mail: Jean-Philippe.Goldman@unige.ch

Criteria in Solid Tumours) [1] in order to build a standardized research database, and to provide a reliable decision support tool.

The results of the assessment follow the RECIST 1.1 classification with these 4 categories: Complete Response (CR), Partial Response (PR), Stable Disease (SD), Progressive Disease (PD). We also aim at classifying the same data in two classes P/NP: P = progressive tumor (PD), and NP = non-progressive tumor (CR, PR or SD).

2. Material and Methods

2.1. Data Extraction

The data used for the training of the automatic classifier come from various places and stages of the SPO project. In an early exploratory step of the project, both oncology teams of CHUV and HUG were asked to manually extract approximately 120 radiology reports, including 6 major tumor cancer types (Breast, Abdominal, Lung, Prostate, Melanoma, Blastoma), preferably with a balanced distribution of cases across the 4 RECIST classes and across the different types of cancer Several radiology modalities for assessment were included (MRI, CT-SCAN, PET-CT). HUG extraction led to 122 radiology reports (HUG122), while CHUV extracted 118 reports (CHUV118). The conclusion of these reports have 31 words in average (min=5, max=54). Here are 3 short examples: 1) Majoration de la condensation lobaire inférieure gauche, 2) Stabilité de la maladie tumorale, 3) Pas de récidive tumorale locale ou à distance

At a second stage of the project, a cohort of patients with a BRAF gene mutation was selected for other longitudinal studies, leading to a group of 108 HUG unique patients representing 303 radiology reports (BRAF303).

The reports were de-identified before processing in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule's De-Identification Standard. The study was approved by SwissEthics (2020-00347).

2.2. Data Annotation

The three sub-corpuses (HUG122, CHUV118, BRAF303) were annotated the same way by 2 to 3 expert annotators from the respective institutions (HUG for HUG122 and BRAF303, and CHUV for CHUC118) on the basis of the conclusion section of the reports only and using the guidelines defined by oncologists involved in the SPO project, and according to the RECIST 1.1 criteria. Additionally, the annotators were asked to add two information with a yes/no label:

- *dissociated response*, for reports with at least two different RECIST classes.
- *low confidence*, when doubt or uncertainty was expressed in the report.

After the annotation of HUG122 corpus by 3 experts, the Cohen's kappa inter-annotator agreement (IAA) was calculated for each pairs of annotators (0.80, 0.79, 0.93) as well as the global IAA (0.84). The computed IAAs of the two other corpus (CHUV118 and BRAF303) were 0.88 and 0.83 respectively.

Eventually, all of the parallel annotations were reviewed by a single expert annotator from HUG, who solved annotation disagreements, yielding to a gold-standard reference corpus of 543 annotated report conclusions with RECIST classification, as described in Table 1, as well as *dissociated response*, and *low confidence* information. The two-class annotation P (progressive) vs. NP (non-progressive) was derived from the RECIST annotation with the label P for PD conclusions, and the label NP for the others.

Table 1. Number of conclusions per RECIST class and per sub-corpus

RECIST	HUG122	CHUV118	BRAF303	Total
Source	HUG	CHUV	HUG	
1 = CR	22	12	93	127
2 = PR	21	30	33	84
3 = SD	36	20	52	108
4 = PD	40	56	110	206
5 = Unknown	3	0	15	18
Total	122	118	303	543

The 18 conclusions labelled as 5 (equivalent to *unknown* in the RECIST classification) are left out for this study. Most of these are conclusions of radiology reports that are not depicting the evolution of an oncology disease, such as the initial assessment. Among the remaining 525 conclusions (hereafter named as *complete dataset*), there are 19 conclusions annotated as *dissociated response*, while the conclusions flagged as low confidence represent 21 items, with 3 conclusions in common (i.e. annotated both as *dissociated response* and *low confidence*). All in all, there are 488 *non-ambiguous* conclusions. This latter set of 488 conclusions is named *filtered dataset*.

2.3. Automatic Classification

The complete dataset was preprocessed with the usual steps of NLP pipelines:

- *character normalization,* (but diacritics were kept) ;
- *removal of stop-words,* (the words implying *negation* were excluded from the stop-word list. In other words, these *negation* words were kept in the resulting preprocessed dataset in order to keep the valence of a phrase to have a clear distinction between *progression vs. no progression*) ;
- *lemmatization,* (using the well-known NLP framework SpaCy [2] together with *fr_core_news_md*, a language model trained on news in French) ;
- use of *CountVectorizer* (a procedure that converts a collection of text documents to a matrix of token counts. Some preliminaries attempts showed the best results with counts combining words, 2-grams and 3-grams of words).

Four techniques from traditional machine learning (ML) domains [3] and two deep learning (DL) techniques were selected and trained with the data.

1. SVM-RBF: A support vector machines (SVM) technique is a supervised learning method mainly used for classification, outliers detection and regression. It has the advantage of being effective in high dimensional spaces. Also it uses a subset of training points in the decision function (called support vectors), for memory efficiency purpose. It allows various kernel function for the decision function like for example the Radial Basis Function (RBF).
2. SVM-Lin: The Linear SVM is a similar technique using a linear decision function. The Stochastic Gradient Descent (SGD) approach is using a convex loss function, which was proven to have good results with large-scale and sparse machine learning problems found in NLP.

3. NB: The Naive Bayes method is also a supervised learning algorithm based on the Bayes' theorem with the naive assumption of conditional independence between every pair of features given the value of the class variable. [4]
4. LR: the Logistic Regression is a multiclass classifier using the one-vs-rest scheme, also known as logit or Maximum Entropy regression. [5]

The two deep learning techniques used in this study were:

1. Feed-forward neural network (FNN): the FNN includes an embedding layer (in our configuration, with embedding bag of size 128), a mean over vectors and a linear layer (of size 128) outputting an array of size *num_class*. *num_class* represents the number of classes (2 for the P-NP task, 4 for the RECIST task in our case). The FNN was trained with a CrossEntropy Loss function with a learning rate of 10^{-2}. To avoid overfitting, an early stopping strategy (ESS) was used to stop the training as soon as the validation accuracy reached a ceiling.
2. Convolutional network (CNN): the CNN also includes an embedding layer, several layers of Conv-2D (4 layers in our configuration with kernel sizes of (2,128) to (5,128)), a linear layer (of size 512, and an output array of size *num_class*) and a Softmax layer. It was also trained with a CrossEntropy Loss function with the ESS, with a batch size of 16, and a learning rate of 10^{-5}.

Both NN trainings used 4-fold cross-validation (i.e. the test is 25% of the dataset). During the training, 20% of the training set is used as validation set.

3. Results

The data were used to train the 6 classifiers for the targeted categories (RECIST and P-NP) with and without filtering out the conclusions labeled as *dissociated response* or *low confidence*. The *complete dataset* (i.e. non-filtered) included 525 conclusions while the *filtered dataset* included 488 conclusions. The following table shows the accuracy obtained for all the configurations.

Table 2. Accuracy of the 6 learning techniques for complete and filtered dataset and for the two classification tasks (RECIST, PNP)

	complete dataset (n=525)		filtered dataset (n=488)	
	RECIST	PNP	RECIST	PNP
SVM-RBF	0.74	0.87	0.77	0.88
NB	0.75	0.83	0.76	0.84
SVM-Lin	0.79	0.85	0.80	0.87
LR	**0.82**	**0.88**	**0.82**	**0.90**
FNN	0.77	0.85	0.81	0.89
CNN	0.78	**0.88**	0.80	0.89

Table 2. shows similar results across all 6 techniques, ranging from 0.74 for the RECIST task with the complete dataset and NB approach, to 0.82 for LR. Comparing with inter-annotator agreement as mentioned in section 2.2, LR is also the best choice for the 2-class PNP task (with an accuracy of 0.88). As for the *filtered dataset* (n=488), the removal of the 18 ambiguous conclusions yielded slightly better results as expected with an accuracy of 0.82 (respectively 0.90) for the RECIST task (resp. for the PNP task). Figure 1. shows the confusion matrix for the LR approach with the well-classified

conclusions on the diagonal and the erroneous classifications (18 items among 102) evenly distribute. Some surprising cases show a mix up between radically opposite labels (Complete Response vs. Progressive Disease). Additional qualitative introspection should be done to diagnose these wrong classifications.

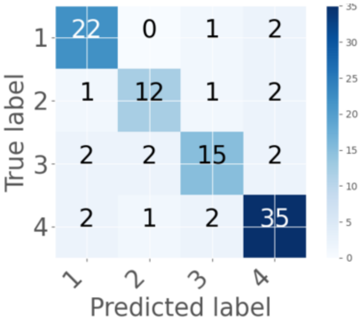

Figure 1. Average confusion matrix for the validation set (20% of the complete dataset) with the LR approach for the RECIST classification tasks (1=CR, 2=PR, 3=SD, 4=PD)

4. Discussion

The two main results of this study are the followings. First, the resulting accuracy is always comparable to the human inter-annotator agreement. Second, across the variety of ML approaches tests in this study, Logistic Regression has the best results for all configurations (*complete* vs *filtered dataset*, and RECIST vs PNP). Also, one can notice that the DL approaches (FNN and CNN) while performing worse than the LR approach, have comparable results with the other ML approaches. The surprisingly equivalent performances observed with DL techniques can possibly be explained by the small size of the corpus. Using the results, several other studies could be designed. The confusion matrix suggests to implement some other target task such as regression or ordinal classification [6] (instead of classification), to take in account the hypothetic proximity of consecutive classes (i.e. CR (1) conclusions are more alike PR (2) than PD (4)).

Additionally, using a larger dataset, the accuracy variation across institutes (CHUV vs. HUG) could be tested. Finally, one should try to reproduce the classification task at the sentence level instead of whole conclusions as the latter can describe different evolutions on various tumor sites and get better accuracy scores.

Acknowledgements: This work was funded by the Swiss Personalized Health Network – SPHN.

References

[1] Eisenhauer EA, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009 Jan;45(2):228-47.

[2] Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

[3] Buitinck et al. (2013). API design for machine learning software: experiences from the scikit-learn project

[4] H. Zhang (2004). The optimality of Naive Bayes. Proc. FLAIRS.

[5] C.M. Bishop (2006) Pattern Recognition and Machine Learning, Chapter 4.3.4

[6] Frank, E., & Hall, M.A. (2001). A Simple Approach to Ordinal Classification. ECML.