Artificial Intelligence In Medicine 91 (2018) 72–81

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Automatic classification of radiological reports for clinical care

Alfonso Emilio Gerevini[a], Alberto Lavelli[b,*], Alessandro Maffi[a], Roberto Maroldi[a,c], Anne-Lyse Minard[d], Ivan Serina[a], Guido Squassina[c]

[a] Università degli Studi di Brescia, Italy
[b] Fondazione Bruno Kessler, Italy
[c] Spedali Civili di Brescia, Italy
[d] Univ Rennes, Inria, CNRS, IRISA, France

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Text mining
Text classification
Analysis of radiological reports
Machine learning for information extraction
NLP for biomedical texts

ABSTRACT

Radiological reporting generates a large amount of free-text clinical narratives, a potentially valuable source of information for improving clinical care and supporting research. The use of automatic techniques to analyze such reports is necessary to make their content effectively available to radiologists in an aggregated form. In this paper we focus on the classification of chest computed tomography reports according to a classification schema proposed for this task by radiologists of the Italian hospital *ASST Spedali Civili di Brescia*. The proposed system is built exploiting a training data set containing reports annotated by radiologists. Each report is classified according to the schema developed by radiologists and textual evidences are marked in the report. The annotations are then used to train different machine learning based classifiers. We present in this paper a method based on a cascade of classifiers which make use of a set of syntactic and semantic features. The resulting system is a novel hierarchical classification system for the given task, that we have experimentally evaluated.

1. Introduction

The use of electronic health record (EHR) in the last years has allowed hospitals to collect a large amount of digital contents (both structured data and narrative text). Such contents have generated new challenges and opportunities in the medical domain since, for example, they can be used to improve the clinical workflows, the efficacy and quality of patient care and can also be used in research in medicine. In particular, natural language processing (NLP) techniques are fundamental and efficient for the automatic extraction of information and allow an effective use of unstructured clinical narratives of the EHR, including radiological reports.

In our context, the use of automatic techniques to analyze clinical narratives (e.g., radiological reports) is fundamental in order to make their content effectively available to radiologists in a structured form; in fact, around 5500 reports of chest computed tomography are generated every year by the radiology department involved in this project, and all these unstructured data cannot be easily summarized and evaluated by humans.

In this paper we focus on the automatic classification of chest computed tomography reports according to a template proposed by the radiologists of the University Hospital of Brescia (*ASST Spedali Civili di Brescia*). The potential advantages of a reliable automatic classification of both old and new reports invest into diverse areas. As neoplastic and non-neoplastic imaging studies are distinguished, the automatic classification will fill a database to be used for retrieving cases for research or teaching purposes. Other relevant benefits concern areas such as logistics and health care management. As the automatic classification of reports enables to separate positive from negative imaging studies or first imaging studies from follow-ups, it could be used to know how often the provisional diagnosis is confirmed by a positive imaging study and to monitor the frequency of follow-up examinations requested by different physicians.

Our goal is to build a system that can be used to automatically classify all the reports generated until now. Moreover, it could be integrated in the software used by radiologists for writing the reports; this would allow to obtain a “real time” classification of a report (as soon as the radiologist has written it), which should then be confirmed (or modified, if needed) by radiologists. This would produce a twofold effect: (i) the manual validation of the automatic results of our system would help to build a more accurate classifier; (ii) if accurate enough, this would reduce the classification effort required by the physician.

We address an application involving standard NLP and machine learning (ML) techniques in the medical (radiological) domain. The novelty of the work is along the following lines: the type of classification is different and more complex than the ones of previous

* Corresponding author.
E-mail addresses: alfonso.gerevini@unibs.it (A.E. Gerevini), lavelli@fbk.eu (A. Lavelli).

https://doi.org/10.1016/j.artmed.2018.05.006
Received 2 November 2017; Received in revised form 10 May 2018; Accepted 25 May 2018

applications in the radiological domain; the hierarchical classification schema we describe has been designed by expert radiologists and it is here proposed as a reference schema for other radiology departments; the schema poses additional difficulties with respect to a "flat" classification (e.g., the need of aggregating the outcomes of the classification at different levels, carried on using rules and knowledge elicited from expert radiologists); the experimental comparison of alternative techniques for our classification task (three methods and several state-of-the-art ML algorithms).

The main difficulties and challenges of our radiological report classification are: (i) the use of only texts and no images (incompleteness of the information); (ii) given that the patients can have direct access to the reports, sometimes the radiologists intentionally provide fuzzy/ambiguous explanations; (iii) the multiplicity of the information (different body parts, different sites, etc.).

We organize the paper as follows. In Section 2 we define some background concepts and review related work. Then in Section 3 we describe our classification schema and our dataset. In Section 4 we give a detailed description of our classification system and in Section 5 we present the evaluation of our approach. Finally, in Section 6 we conclude the paper with future work.

2. Background and related work

A radiology report is the formal product of a diagnostic imaging referral, used for communication and documentation purposes. In general there are different guidelines for effective reporting of diagnostic imaging, although essentially a report consists of free text, possibly organized in a number of standard sections. Medical reports and clinical narratives are characterized by non-standard language: they contain abbreviations, ungrammatical language, acronyms and typing errors; this is due to the fact that reports are often written in haste or dictated to speech recognition software. In addition, abbreviations and acronyms are sometimes idiosyncratic to the specific hospital or department.

Natural language processing techniques are needed to convert the unstructured text of these reports into a structured form, therefore enabling automatic identification of information. NLP applications rely on a sequence of steps that extract structured textual features from the radiology report. Usually, the first step is *segmentation*, i.e. splitting the reports into their sections. The following steps can then operate on a subset of the sections or apply a specific weight on the content of different sections. The text is then divided into sentences (*sentence splitting*) and the sentences into tokens (*tokenization*). Additional normalization steps can follow at the token level, such as determining the lexical root of words (*stemming*), fixing spelling mistakes or expanding abbreviations to their full form. Through morpho-syntactic analysis it is then possible to determine the *part of speech* of words (e.g. noun, verb, adjective) and their grammatical structure (e.g. noun phrase, verb phrase, prepositional phrase). These steps enable to perform semantic analysis, i.e. assigning meaning to the words and phrases by linking them to semantic types and concepts (*concept recognition*). A further step is *negation detection*, i.e. checking whether concepts or relations in the text are negated. The final result of these steps is a set of features that can be used for the actual task, for instance, text classification. The features can be processed by an automatically generated classifier (machine learning approach) or by a set of rules hand-crafted by experts. Hybrid approaches are also possible.

In literature we find three main application areas related to automatic analysis of free text:

- *text classification* (or categorization) is the task of deciding, given a text and a predefined set of classes, which class the text belongs to;
- *information extraction* is the process of acquiring information by analyzing a text and extracting occurrences of specific entities and of relationships among objects;
- *information retrieval* is the task of finding documents that are relevant for the user's information needs.

Pons et al. [1] present a systematic review of NLP applications for radiology developed until 2016, both in operational use or not. They point out five main categories of study: (a) diagnostic surveillance, (b) cohort building for epidemiological studies, (c) query-based case retrieval, (d) quality assessment of radiologic practice, (e) clinical support services. Most of the applications studied in [1] are developed for English and would require substantial changes to work for Italian. We can in particular mention the work presented in [2] about recognition of recommendations in radiology reports and a work on the classification of radiology reports in two classes: whether a report contains a "cancer alert" or not [3]. The latter reported a *F*-measure of 0.77 on the binary classification. Khachidze et al. [4] have worked on the classification of clinical records written in the Georgian language. The relevant point of their work for us is the fact that they performed a multi-level classification. They defined two levels: an up-level which consists in determining which clinical exam has been performed (ultrasonography, endoscopy or X-ray), and a second level which concerns the site of the exam. The possible classes for the second level depend on the classification of the up-level: if the report is about an X-ray then the possible sites are chest, abdomen, etc.; if it concerns an ultrasonography then the site could be for example liver, biliary system, etc.; if it is an endoscopy, no site has to be specified. They propose a method based on support vector machines and k-nearest neighbors using 13,716 reports as training set and 11,140 as test set. Our classification system has more levels than the work in [4], it addresses a different task and works for a different language; moreover, our current system is based on a much smaller training data set. For the first level they obtained an F_1 between 0.87 and 0.91 and for the second level an F_1 between 0.50 and 0.93.

In a recent paper, Yim et al. [5] propose an approach to the problem of automatically identifying tumor-related information in radiology reports based on a structure called *event* (or template). A tumor event is defined as a predefined set of related concepts that encodes information related to a predetermined event representative, such as "lesion" or "2.4 cm". An attribute is a field in the event structure that takes a range of values that may or may not be bounded by a closed set of variables. The work focuses on classification for three specific tumor event attributes: negation (i.e., present, absent), temporal (i.e., past, current), and malignancy (i.e., benign, indeterminate, malignant, unknown). Differently from our approach, they do not perform a multi-level analysis, and the attributes are slightly different from ours; so a direct comparison can be difficult. Our approach is more structured, involving a higher number of levels and classes, as requested in the context of our classification task. The performance in terms of F_1 is the following: negation identification: 0.94; temporality classification: 0.62; malignancy classification: 0.77.

As for Italian, some works which use supervised learning in order to extract information from radiology reports are [6–8]. In [7], a corpus of manually annotated reports was created to be used as a training set. Segments of text were annotated with tags representing concepts of interest in the radiological domain. Using the same dataset of radiology reports, Marcheggiani and Sebastiani [8] tested the impact of training data quality on the accuracy of information extraction systems as applied to the clinical domain. In [6] the aim is to find relations among biomedical entities, i.e. addressing an application task different from ours. A very large set of reports was automatically annotated with NER (Named Entity Recognition) [9] tools to be used as a training set. To automatically extract medical entities, standard taxonomies (e.g. Snomed-CT, ICD9) can be used; in some cases, entities can then be mapped to their unique UMLS CUI (Concept Unique Identifier) [10].

Our work consists in the classification of radiology reports following a new multi-level schema designed with domain experts for chess computed tomography, whereas the previously mentioned works focused mainly on a unique level or on 2 levels [4]. Due to the multi-level

aspect of our classification schema we perform the classification through a cascade of classifiers. They are used to annotate a report at the sentence-level and to classify it in different classes.

Given the differences between our classification task and the related work mentioned above, a direct comparison of the relative classification performance is difficult and does not seem very informative.

2.1. Machine learning algorithms

Machine learning is a field of artificial intelligence addressing the question of how to construct computer programs that automatically improve their performance with experience, that in supervised learning is specified as training data sets. Due to the effectiveness of the existing machine learning techniques and the always increasing amount of available (structured and unstructured) data, in the last decade the number of research projects and practical applications that use machine learning has dramatically increased, covering a wide spectrum of practical tasks.

For the implementation and evaluation of our system for the reports classification, we focused on five (supervised) machine learning techniques that have been successfully applied in different text mining applications described in the literature: Naive Bayes classifiers [11], decision trees [12], random decision forests [13,14], neural networks [15] and support vector machines [16].

The Naive Bayes classifier [11] is a predictive machine-learning method based on the Bayes rule of conditional probability, which is known for creating simple but well performing models in the field of document classification. It makes use of all the attributes (features) characterizing the data, and analyses them individually as though they are equally important and *independent* of each other. Although this is a strong assumption, it usually allows to build simple models that sometimes work surprisingly well.

A support vector machine (SVM) [17] is a learning method that can make use of specific non-linear functions called *kernels*, in order to automatically translate the instances of the training data in a multi-dimensional space where linear classification techniques can be directly applied. The classification model of a SVM is formally defined by a separating hyperplane, in the new multi-dimensional space, that maximally separates the binary labeled training data. Even if the classification of the data is not binary, a SVM handles it as if it were binary, and it completes the analysis through a series of binary assessments on the data. SVMs can work surprisingly well also with a low number of observations across many data attributes. SVMs have been widely used for text categorization due to their robustness, ability to generalize well in high dimensional feature spaces, and reduced effort for feature selection that makes their application to text classification considerably easier than other methods where selection of the data attributes is more crucial.

Decision Trees [12] use one of the most popular and often effective learning methods in data mining for solving classification tasks. The internal nodes of a decision tree denote the different data attributes; the branches between nodes correspond to possible value sets of these attributes in the observed samples; while the terminal nodes indicate the classification value of the target (or class) variable. A key advantage of decision trees with regard to other approaches is that the classification model implemented by a decision tree can be easily understood by humans, which is very important in the medical field to make the doctors better trust the automatic classification results. On the other hand, the effectiveness of decision trees as a general tool to generate classification models is affected by the reduced space of hypotheses for the target variable that decision trees can represent.

Random decision forests [13] is an ensemble learning method [18] that constructs a number of decision trees at training time, and define a classification variable that is the mode of the class variables of the individual trees. For constructing each individual tree of the random forest, a randomly chosen subset of the data attributes is used. Random decision trees can provide better results than a single decision tree, although the final classification model is more difficult to understand for humans.

Finally, neural networks [15] are a well-studied method that can be very effective in complex contexts. Neural networks are suited when the classification model of the target variable can be (highly) non-linear and complex, the training data may have hidden (unseen) relationship that are inferred by network training, and the amount of training data can be very large and noisy. The classification models that we built using neural networks are based on feed-forward networks trained by the well-known back propagation algorithm. On the other hand, the classification behavior of neural networks is hard to understand for humans.

3. Data representation and annotation

The proposed system for reports classification is based on a classification schema that we defined in strict collaboration with the radiologists of *Spedali Civili di Brescia*. The schema consists of five high level classes that may assume two or more different values. Our approach relies on supervised machine learning methods, so it was necessary to perform a manual classification of a set of reports to train models and to evaluate them.

In this section, we describe the classification schema, then the data used and their manual annotation.

3.1. Classification schema

Fig. 1 shows the classification schema that we designed with the radiologists. It is composed of five levels:

1. examination type (*first examination* or *follow-up*);
2. result of the examination (*positive* or *negative*, *stable* or *progressive relapse*);
3. neoplastic nature of the lesion (*neoplastic*, *non-neoplastic* or *lesion with an uncertain nature*);
4. site of the lesion (*lung*, *pleura* or *mediastinum*);
5. type of the lesion (*infectious*, *aspecific* or *uncertain*,[1] *primary*, *metastasis* or *uncertain*).

3.2. Data

We used CT (computed tomography) reports from one of the radiology departments of *Spedali Civili di Brescia*. The reports were extracted from the database of the hospital and anonymized by removing patient names as well as medical staff names. The reports at our disposal are composed of text only, no CT images are associated to the text. The reports may contain three parts: *quesito clinico*, i.e. the reason why the examination is requested by the doctors (often contains a provisional diagnosis made by the doctors using the results of the previous analysis); *quadro clinico*, i.e. the case history of the patient; *referto*, i.e. the report written by the doctor when analyzing the CT. Since the first two elements are not always available in a report, we did not take them into account for our classification task, considering only the text of the reports. We had at our disposal around 10,000 unlabeled reports of chest CT done in 2015 and 2016.

3.2.1. Manual annotation

The manual annotation of a set of reports is necessary in order to develop supervised machine learning methods and to evaluate them.

[1] It was often difficult even for an expert to determine the type of the lesion reading the textual report only. For this reason, we have added the value *uncertain* to level 5. *Uncertain* has to be selected by the expert if not enough information is available in the report about the lesion type.

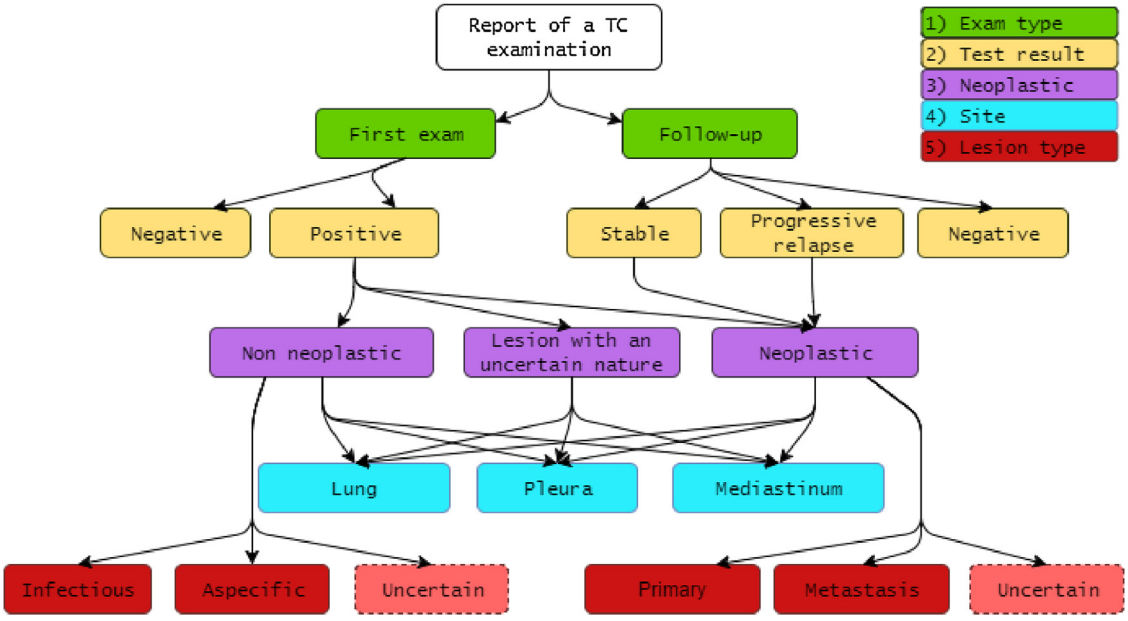

Fig. 1. Classification schema.

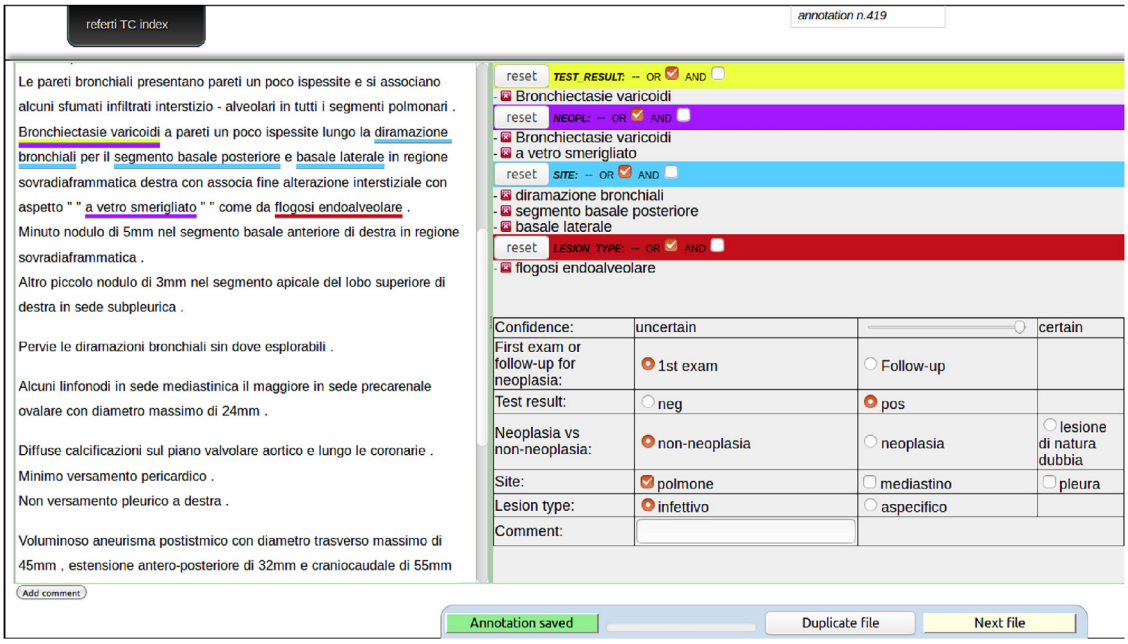

Fig. 2. Web interface used to manually annotate and classify the reports.

To perform the manual annotation, we used the AnnotatorPro tool [19] that we have adapted to our annotation task. AnnotatorPro, built on top of the MT-EQuAl tool [20], is a web interface based on PHP and MySQL. Fig. 2 shows a screenshot of the interface. For each report the annotator performs two tasks: classification of the report through a form (according to the schema described in Section 3.1), and annotation of evidences of the classification in the text (using five tags corresponding to the five levels of the annotation schema). For example, "posterior basal segment" is an evidence for the "Site" *lung*. In Fig. 2, one can see on the left part the radiology report, on the top of the right part the text segments annotated, and on the bottom of the right part the form to fill. The guidelines provided to the annotators are very simple and consist mainly in asking to annotate the maximum extent of text that is an evidence of a level of the classification.

As agreed with the radiologists, the classification schema forces to choose only one test result for each report but reports can describe the evolution of multiple lesions, which can be of different nature or type, and can be located in different sites. For example, a report could describe both a primary neoplastic lesion in the lungs and a nodule in the mediastinum which could be suspected for being a metastasis. Indeed in each report there is a predominant observation or result, which gives the "overall" classification of the report. In this case some heuristic rules, better described in Section 4.3, are used by the radiologists in order to compute the overall classification. For instance, if a neoplastic lesion has been identified on one site and non-neoplastic lesions have been found on other sites, then the report should be classified as "neoplastic" and not as "non-neoplastic".[2] As a consequence, the

[2] As described later in more detail, such rules elicited from the radiologists are implemented in the report classification system.

Table 1
Statistics about the annotated data.

Reports of a chest CT examination: 346						
Exam type	First exam 136			Follow-up 210		
Result	Negative 16	Positive 120		Negative 77	Stable 79	Progressive relapse 54
Site	Lung 188			Pleura 19	Mediastinum 40	
Neoplastic	Non-neoplastic 54			Uncertain 29	Neoplastic 161	
Lesion type	Infectious 25	Aspecific 28	Uncertain 1	Metastasis 53	Neoplastic primitive 28	Uncertain 80

interface enables the annotator to "duplicate" the annotation window for a report (i.e. to annotate more than once the same report) so that every described lesion can be classified.

The annotation task was performed by an expert (a specialist in radiology). Periodically, after he had annotated a certain amount of reports, his classification/annotation was discussed within the project team and, in case of inconsistencies/incorrectness or missing evidences with respect to the adopted classification schema, the manual classification was corrected with the specialist. The amount of these corrections was quite limited and related essentially to some missing annotations, that were not fundamental for the classification of the report, but that are useful as training elements of our classification system.

3.2.2. Data statistics

In total the medical expert annotated 346 reports selected manually out of the 10,000 unlabeled reports in order to represent adequately all the classes of our schema, which have been used to train and test our approach.[3] Reports are composed on average of 174 words. Some statistics about class distribution are presented in Table 1. We can observe that the distribution of values in some classes is unbalanced, i.e. one value is much more frequent than the others. For example, there are very few reports associated to the "Site" *pleura* compared to the "Site" *lung*.

4. Report classification

We developed three versions of our report classification system that use different approaches adopted in information extraction and text classification. The first (Method 1) uses a combination of both information extraction and text classification techniques; the second (Method 2) is based on information extraction techniques only; the third (Method 3) uses text classification techniques only. In the following, first we present the initial text processing step (common to all the methods); then we describe the three classification methods, that are schematized in Fig. 3.

4.1. Text processing

Reports are preprocessed using the TextPro suite [21] in order to extract features from text. TextPro performs sentence splitting, tokenization, different linguistic analysis which gives us morphological analysis, lemmatization, Part-of-Speech tagging, identification of syntactic phrases and time expression detection. Making use of these linguistic modules and of some external resources, a tool developed specifically for this project identifies prefixes and suffixes contained in the words (e.g. *-tomia* [-tomy]), negation cues (e.g. *non* [not]) and the presence of numbers and measurements. In addition, if a word is derived from another it is associated to its derived term, and if it has synonyms a unique (preferred) term is added.[4] For example, the word *diminuzione* [diminution] is associated to two other words: *diminuire* [decrease] (its derived word) and *riduzione* [reduce] (its preferred term). The preprocessing step is represented as process #1 in Fig. 3.

4.2. Automatic annotation

The classification can be improved by automatically identifying sequences of words which are evidences of the different classes. These sequences will be then used as additional features for the classification. In order to annotate the relevant sequences of words, we implemented a supervised machine learning based system. We used the conditional random field (CRF) algorithm through the CRF++ toolkit.[5] For each level of the classification schema a model is built. Different types of features are used: surface features (token types), syntactic features (Part-of-Speech, chunk phrases, negation) and semantic features (tokens, prefixes, suffixes, numerical expressions, lemmas, synonyms). The training data are obtained from the manually annotated reports (see Section 3.2.1). In Fig. 3, the automatic annotation is represented by process #2.

The automatic annotation step cannot be evaluated independently from the classification task as the manual annotation of the reports is partial, i.e. not all the phrases related to one class (e.g. "Site", "Lesion type") have been annotated by the expert, but only those that are evidences of the classification he chose.

This aspect will be further discussed in Section 5.4.3.

4.3. Classification

In the following, for each analyzed method we describe how the final system classifies a new (unseen) report and how it is trained (using the corpus of annotated reports).

4.3.1. Methods 1 and 2

These methods try to distinguish between *first examination* and *follow-up* (first level of the classification schema) in the same way, i.e. searching for typical expressions (patterns) associated to follow-ups in the whole text. The patterns are automatically generated from the training set, using the manual annotations.

[3] The use of all unlabeled reports would have required too much annotation effort for the domain expert.

[4] The list of synonyms and their preferred terms has been built manually by the authors and contains 173 words.

[5] https://taku910.github.io/crfpp/.

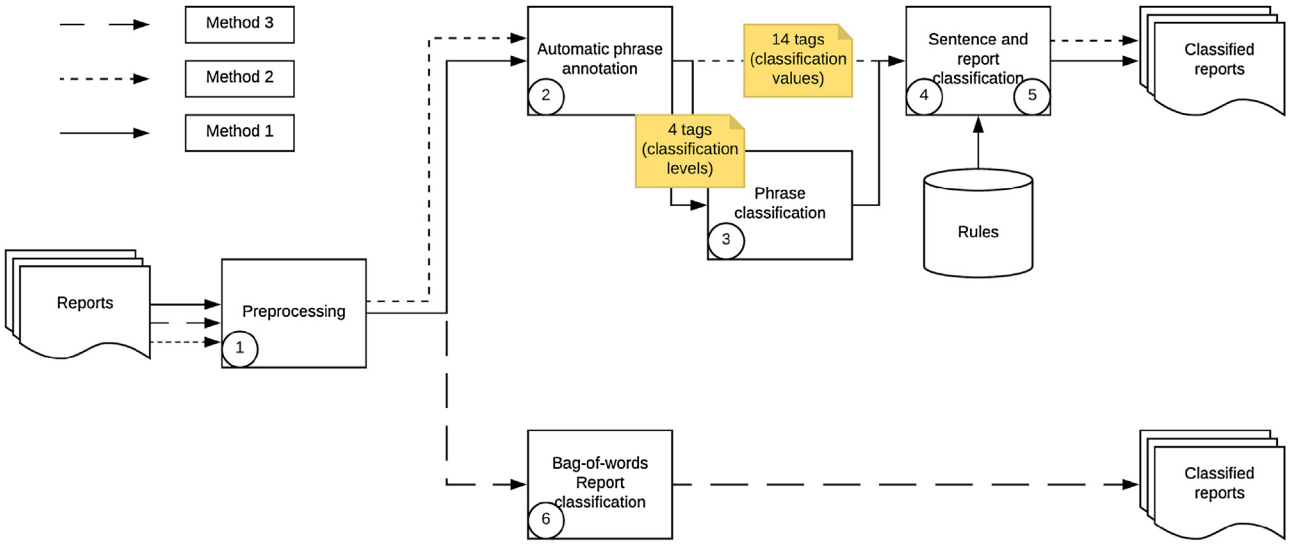

Fig. 3. Schema of the three proposed classification methods.

Then the text is divided into sentences, which are automatically annotated (with the tool described in Section 4.2) and independently classified according to the remaining four levels of the classification schema (process #4 in Fig. 3). A sentence is associated to a value of a certain level if it contains an automatically annotated evidence of this value. For example, if in a sentence "increase of lesion size" is annotated as *progressive relapse*, then the whole sentence is classified as *progressive relapse*.

The final classification of a report is obtained by merging together all the sentence classifications (process #5 in Fig. 3), according to the following heuristic classification rules from the medical domain that are used by the radiologists:

- if the report is a follow-up and there is at least one sentence classified as *progressive relapse*, then this classification prevails on *stable*;
- if at least one sentence is classified as *neoplastic*, then the report is classified as *neoplastic* (*neoplastic* prevails on *lesion with an uncertain nature* which prevails on *non-neoplastic*);
- the sites are collected from all positive sentences;
- if the report is associated to *neoplastic* in the third classification level and there is at least one sentence classified as *metastasis* for the fifth classification level, then *metastasis* prevails on *primary*. By default (no sentences are classified as *metastasis* or *primary*) the lesion type is classified as *uncertain*;
- if the report is associated to *non-neoplastic* in the third classification level and there is at least one sentence classified as *infectious* for the fifth level, then *infectious* prevails on *aspecific*. By default (no sentences are classified as *infectious* or *aspecific*) the lesion type is classified as *uncertain*.

Method 1 and Method 2 differ in the way the automatic annotation of the sentences is performed. In Method 1, automatic annotation follows two steps:

- The automatic annotation tool classifies significant words or phrases according to four different tags. The tags are identical to the ones used by the annotator: each one is associated with a level of classification.
- The tagged words and phrases are classified into a specific class value in order to perform sentence classification (for example, an item tagged with tag "Site" must be classified into *lung*, *pleura* or *mediastinum*) using the bag-of-words model. In this model, a text is represented as the bag of its words, disregarding syntactic structure and even word order but keeping multiplicity and the frequency of occurrences of each word is used as a feature for training a classifier. Eight different text classifiers are trained for this purpose, using the annotated sections of the reports.

On the other hand, Method 2 relies on an enhanced version of the automatic annotation tool, which tags the text using one specific tag for each class value (e.g. for the "Site" there will be three tags, one for *lung*, one for *pleura* and one for *mediastinum*), instead of using four generic tags. A second step (i.e. classifying tagged words and phrases into a specific class value) is not needed, since the 14 different tags corresponding to all possible class values (except *uncertain*) of the classification scheme are used in this case.

We now give an example of automatic annotation and classification using Method 1. Let us consider a report containing the following two sentences (the most significant of the full report)[6]

"… Increase in size of the parenchymal lesion in the left upper lobe apicoposterior segment. The hilar and mediastinal adenopathies remain unchanged …"

- …" < TEST_RESULT > Increase in size < /TEST_RESULT > of < NEOPL > the parenchymal lesion < /NEOPL > in < SITE > the left upper lobe apicoposterior segment < /SITE > . < SITE > The hilar and mediastinal < /SITE > < NEOPL > adenopathies < /NEOPL > remain [TEST_RESULT] unchanged [/TEST_RESULT]…"
- …" < TEST_RESULT_PROGRESSIVE_RELAPS > Increase in size < / TEST_RESULT_PROGRESSIVE_RELAPS > of < NEOPL_NEOPLASTIC > the parenchymal lesion < /NEOPL_NEOPLASTIC > in < SITE_LUNG > the left upper lobe apicoposterior segment < / SITE_LUNG > . < SITE_MEDIASTINUM > The hilar and mediastinal < /SITE_MEDIASTINUM > < NEOPL_NEOPLASTIC > adeno pathies < /NEOPL_NEOPLASTIC > remain < TEST_RESULT_ STABLE > unchanged < /TEST_RESULT_STABLE > …"
- The first sentence is classified as *progressive relapse–neoplastic–lung*, w.r.t. test result (level 2 of classification), nature of the lesion (level 3) and site (level 4), respectively; the second sentence is classified as *stable–neoplastic–mediastinum*, w.r.t. the same three levels.
- By applying the rules described in Section 4.3.1, we obtain the following overall report classification:

[6] The actual report is in Italian but here we have translated the sentences into English for better readability.

Table 2
Accuracy and macro-averaged *F*-measure using 10-fold cross-validation for classifying manually tagged words and phrases into specific class values, on the labeled data set considering different machine learning algorithms. In bold the best result of each row w.r.t. the *F*-measure.

	NB		SMO		J48		RF		MLP	
	Acc	FM	Acc	FM	Acc	FM	Acc	FM	Acc	FM
`Result`	97.2	**96.3**	96.9	95.9	97.2	96.2	96.7	95.5	96.4	95.2
`Result f-up`	87.3	86.9	91.1	**90.9**	87.3	86.7	89.2	88.9	87.2	87.0
`Neoplastic`	82.6	**75.7**	83.8	73.0	81.0	71.5	83.4	74.5	83.4	73.2
`Site lung`	87.4	84.5	93.0	**90.5**	90.4	87.6	92.8	90.1	91.3	88.0
`Site med.`	97.2	**86.5**	97.2	**86.5**	97.2	**86.5**	97.2	**86.5**	97.0	85.7
`Site pleura`	94.3	88.6	96.0	92.2	92.8	85.2	95.7	91.9	96.0	**92.9**
`L.t. neopl.`	91.2	90.3	92.5	**92.1**	80.0	77.6	91.2	90.6	92.5	92.0
`L.t. non neopl.`	87.7	**87.5**	86.0	85.4	68.4	68.3	82.5	81.8	86.0	85.7

Exam type: *follow-up*
Test result: *progressive relapse* (Rule R1)
Nature of the lesion: *neoplastic* (Rule R2)
Site: *lung, mediastinum* (Rule R3)
Lesion type: *uncertain* (no sufficient information to discriminate between primary or metastasis) (Rule R4)

This is what happens with Method 1. Method 2 starts with "phrase classification" and Method 3 directly performs "report classification" (step 4).

4.3.2. Method 3

This method is based on the bag-of-words model (which is commonly used in text categorization) according to which, the whole text of the report is used to obtain the classification (process #6 in Fig. 3). Neither the manually annotated data nor the automatic annotation tool are used in this case. Given the high number of features that can be generated, we used the *information gain attribute ranking* for the selection of the most relevant features [22].

5. Evaluation and discussion

The hierarchical classification system that we propose in this paper combines different NLP and machine learning techniques in order to effectively classify complex textual radiological reports. In this section, we evaluate the performance of our system with the different classification techniques it incorporates.

For applying the bag-of-words model in Step 2 of Method 1 and in Method 3, we experimented different learning algorithms: the Naive Bayes classifier (NB); the Sequential Minimal Optimization algorithm (SMO) for training the support vector machines; the J48 implementation for the decision trees; Random Forests (RF); the MultiLayer Perceptron model (MLP) for training the neural networks. To implement the different text classifiers, we used the Weka open source Java data mining library [23], with the default configurations for each classifier.

5.1. Evaluation metrics

The performance was evaluated using the following measures:

- *accuracy* (Acc), which is the number of correct predictions divided by the total number of predictions.
- *macro-averaged F-measure* (FM), which is obtained computing the *F*-measure locally over each category first and then considering the average value. The *F*-measure is the harmonic mean of *precision* (the number of correct positive results divided by the number of all positive results) and *recall* (the number of correct positive results divided by the number of positive results that should have been returned), i.e.

$$F\text{-measure} = 2 \cdot \frac{1}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Accuracy is a standard and intuitive performance measure used in different research papers; in the following we analyze the proposed results considering primarily the macro-averaged *F*-measure since it balances precision and recall of our classifiers determining a more significant performance index.

5.2. Experimental results

In Table 2, we show the efficacy of machine learning algorithms in classifying tagged words and phrases into specific class values (Step 2 of Method 1). The results are obtained using 10-fold cross validation on the phrases that were manually annotated by the experts. In particular, we used:

- two text classifiers for the "Result" level (`Result` and `Result f-up`), which respectively distinguish between *positive* or *negative* and *stable* or *progressive relapse*;
- one classifier for the "Neoplastic" level (`Neoplastic`) which distinguishes among *neoplastic*, *non neoplastic* or *uncertain*;
- three classifiers for the "Site" level (`Site lung`, `Site pleura` and `Site med.`);
- two classifiers for the "Lesion type" level, one for neoplastic reports (`L.t. neopl.`) which distinguishes between *primitive* or *metastasis* and one for non neoplastic reports (`L.t. non neopl.`) which distinguishes between *infectious* or *aspecific*.

The experimental results in Table 2 show the general good performance of support vector machines using SMO (with average FM equal to 88.3%), that are slightly better than RF (average FM 87.5%) and MLP (average FM 87.5%). NB is slightly worse than RF and MLP (average FM 87.0%) and J48 has the lowest performance (average FM 82.5%).

In Table 3, we analyze the results of the classification on a test set (consisting of about the 20% of the 346 reports) considering the classification Method 1 described in Section 4.3 with different supervised machine learning algorithms in Step 2. Note that in Tables 3 and 4 we have an additional classification category with respect to the categories in Table 2: `Exam type` (first table line), with values *first examination* and *follow-up*. The relative results were obtained using either patterns (Method 1 and Method 2) or a classifier (Method 3). From the results in Table 3 we observe that the Random Forest approach provides in general the best results, but also SMO and MLP provide competitive results. More specifically, the Random Forest approach performs

Table 3
Evaluation of the classification method 1 on the test set in terms of accuracy and macro-averaged *F*-measure. The last column reports the number of documents for the different classes in the test set. In bold the best result of each row w.r.t. the *F*-measure.

	Method 1										Number of reports
	NB		SMO		J48		RF		MLP		
	Acc	FM	Acc	FM	Acc	FM	Acc	FM	Acc	FM	
Exam type	95.6	94.9	95.6	94.9	95.6	94.9	95.6	94.9	95.6	94.9	68
Result	77.9	67.1	79.4	70.1	79.4	70.1	82.4	**75.5**	82.4	**75.5**	68
Result f-up	54.8	67.1	67.7	70.1	67.7	70.1	67.7	**75.5**	64.5	**75.5**	31
Neoplastic	70.8	56.1	77.1	**66.0**	60.4	45.3	75.0	57.0	70.8	56.3	48
Site lung	70.8	52.1	72.9	53.5	72.9	53.5	75.0	**55.0**	70.8	52.1	48
Site pleura	87.5	46.7	87.5	59.1	89.6	**61.5**	89.6	47.3	89.6	47.3	48
Site med.	72.9	**67.9**	72.9	**67.9**	70.8	66.1	72.9	**67.9**	66.7	63.6	48
L. t. neopl.	66.7	**64.3**	63.6	60.8	63.6	60.8	63.6	60.8	63.6	60.8	33
L.t.nonneopl.	70.0	**69.4**	60.0	57.8	60.0	57.8	60.0	57.8	60.0	57.8	10

Table 4
Evaluation of the three classification methods on the test set in terms of accuracy and macro-averaged *F*-measure. The last column reports the number of documents for the different classes in the test set. The results proposed for Method 1 and Method 3 have been obtained using Random Forests (RF). In bold the best result of each row w.r.t. the *F*-measure.

	Method 1		Method 2		Method 3		Number of reports
	Acc	FM	Acc	FM	Acc	FM	
Exam type	95.6	94.9	95.6	94.9	98.5	**98.3**	68
Result	82.4	75.5	80.9	71.5	83.8	**78.0**	68
Result f-up	67.7	**75.5**	74.2	71.5	54.8	54.4	31
Neoplastic	75.0	**57.0**	62.5	32.8	73.5	52.0	48
Site lung	75.0	55.0	70.8	**55.8**	87.8	46.7	48
Site pleura	89.6	47.3	87.5	46.7	93.8	**48.4**	48
Site med.	72.9	67.9	81.3	**70.5**	61.2	42.4	48
L. t. neopl.	63.6	**60.8**	60.6	43.6	69.7	49.3	33
L.t.nonneopl.	60.0	**57.8**	30.0	24.4	63.6	42.2	10

extremely well for the upper levels of the classification scheme with accuracy for the Result category that is more than 3 percentage points better w.r.t. SMO, J48 and NB, and more than 5 percentage points better w.r.t. FM. This behavior is essentially related to a better selection of the relevant features performed by the RF Algorithm, in contrast to the plain J48 Algorithm that uses all the available features. Quite interestingly, we can observe that the NB approach performs better than the other approaches for the lower levels of the classification scheme, accuracy for the L.t.nonneopl. category that is 10 percentage points better w.r.t. the other approaches, and more than 11 percentage points considering FM. This more robust behavior of NB w.r.t. other approaches could be essentially related to the assumption of the independence of the features that distinguishes NB from the other approaches and that could be better satisfied in the lowest levels of the classification scheme.

In Table 4, we analyze the results of the classification on the previously described test set considering the three classification methods described in Section 4.3. For Method 1 and Method 3 we show the results obtained using Random Forests (RF) for building the text classifiers. For Method 3 we omit the results obtained with other machine learning algorithms as they are very similar to each other and, overall, Random Forests performs slightly better. We can observe that Method 1 provides in general the best results; moreover, Method 2 seems quite competitive except for the last level of classification, i.e. L.t. nonneopl., which is probably related to the low number of elements for this class in the training set. Concerning Method 3, which uses a simple bag-of-words approach not requiring the automatic annotation of the text, we can observe that for the most specific levels (e.g. "Site", "Result" for follow-ups) the automatic annotation of evidences in the text improves the classification, while, for the upper levels, the bag-of-words approach obtains better results.

We think that for the most specific levels it is necessary to analyze the reports at sentence level as, for instance, whether a single sentence is negated or not could affect the classification of the report. This can represent a limitation for Method 3 as it is based on the bag-of-word approach, which does not consider syntactic information or word order in the reports. We believe that the significant training effort required for Methods 1 and 2 (i.e. manually annotating the reports as described in Section 3.2.1) can lead to a better in-depth report classification.

Overall, among the three proposed and evaluated methods, the best candidate for use in a real production environment appears to be Method 1, which is the first method we intend to integrate into the software used by the radiologists for writing the reports.

Moreover, we would like to point out that, using Methods 1 and 2, our system is also able to identify the more relevant sentences in the text; this represents an interesting perspective since it could allow the physician to read and understand a report more easily. In the future, we plan to better analyze the advantages of the automatic annotation tool for highlighting the relevant sentences of a report in order to improve both report production and report visualization.

5.3. Error analysis

In order to better understand the behavior of the proposed methods, we performed an error analysis. We observed that one cause of wrong classification is the low recall of the automatic annotation module (see Section 5.4.3 for a discussion on its performance), especially when using Method 2. With Method 1 there are many false positives for the "Lesion type" level. This is due to the fact that the information available in the text of a single report does not always enable the expert to identify the lesion type, whereas the system can find partial information and associates it to a type of lesion. The classification rules for Method 1 and Method 2 are strictly related to the presence of an annotation associated to the "Neoplastic" level in the reports. Indeed, if no text segment has been annotated for the "Neoplastic" level, the sentence is considered as negative. This may cause false negatives when classifying for the "Result" level, as some reports will be classified as *negative* even if some text segments indicate a positive result. With Method 3 we observe that less reports are correctly classified as *progressive relapse* than with Method 1 and 2. This can be explained by the presence of terms indicating stable results in sentences that have not to be

considered in the classification (mainly because not about a neoplastic lesion). With Method 1 and Method 2 only the evidences found in positive sentences will be considered, which enables the system to get a more precise classification.

5.4. Further evaluation

5.4.1. Aggregated evaluation for all levels

In Section 5.2 we have evaluated our methods on each classification level independently; i.e. the correct classification for the upper levels was used while classifying the following levels. In order to have an idea of the performance of the proposed methods in a situation when no manual classification is available for a report, we have performed another evaluation which consists in computing the number of the reports correctly classified for *all* the classification levels. We have computed it for Method 1, using the RF algorithm.

Considering all the 5 levels, 23 reports out of 68 are completely correctly classified (34%). As we have previously mentioned, the fifth level ("Lesion type") is a very difficult one even for the experts. For this reason, we have also computed the number of correctly classified reports only for the 4 top levels and we have obtained 30 reports correctly classified out of 68 (44%). Finally, we have evaluated the first 3 levels as for the department of radiology classifying correctly in the 3 top levels is the most important for their future research. Our system with Method 1 is correctly classifying 37 reports out of 68 (54%). These results show room for improvement. But for a first system dealing with the complex classification task that we addressed, we wanted to assess the performance of the methods at the single level, because we know that errors propagate from one level to another, and good performance at the single levels is very important.

5.4.2. Inter-annotator agreement

We have recently conducted an inter-annotator agreement study for the first three levels of report classification (the most important levels according to the radiologists) asking another expert radiologist to classify all the reports of the test set. For level 1 of the classification schema the agreement was 100%, for level 2 93%, and for level 3 73%. The kappa scores were 1.00, 0.81 and 0.53 respectively. The results of this experiment indicate that the addressed task is well defined and that our approach obtains encouraging performance. A study with the radiologists is ongoing in order to identify the reasons of their disagreement.

5.4.3. Evaluation of the automatic annotation module

As mentioned before, it is difficult to evaluate the automatic annotation step independently from the classification task as the manual annotation of the reports is partial (i.e. not all the phrases related to one class have been annotated by the expert, but only those that are evidences of the classification s/he chose). In order to approximately measure the recall of the automatic annotation module described in Section 4.2, we have conducted an additional experiment. We have asked a radiologist to classify all the reports of the test set considering *only* the automatically annotated sentences which were used by the system to classify the report. We compared the two classifications given by the expert obtained using the whole text and only the annotated part, respectively. The experiment was focused on the third level of the classification schema (examination result). We observed that the agreement between the two classifications was 85%, giving evidence that very often the automatic annotation correctly identifies the parts of the text that are informative for the classification.

6. Conclusions and future work

In this paper we have presented a system for the automatic classification of chest computed tomography reports in Italian. The approach is based on machine learning techniques and relies on a classification schema proposed by the radiologists involved in the project. We have compared the performance obtained by different machine learning techniques. The resulting system is a novel hierarchical classification system showing interesting performance; in fact, the experiments performed on the reports annotated so far show encouraging results.

Afterwards, a sixth additional level has been added to the classification schema. It concerns only the follow-up examination and consists of the origin site of the follow-up, i.e. for which site a follow-up was recommended. In many cases, but not all, the origin site is the same as the site of the lesion. Currently, only part of the corpus is annotated according to such level, but soon we will extend the annotation of the origin site to the whole corpus and take it into consideration in the automatic classification system. In the future we plan to extend the data set considering different radiology departments and different annotators. The experiments described in Sections 5.4.2 and 5.4.3, currently involve 2 radiologists, but we plan to involve more experts soon. Along with these experiments, we are currently collecting new reports annotated by another expert and we plan to include them in the dataset of our experiments in the near future. In this work we focused on reports of chest computed tomography. We plan to extend the classification to other parts of the body (e.g. encephalon) extending consequently the classification schema. We also intend to study the application of additional machine learning and text processing techniques, such as deep neural networks [24] and more sophisticated document representations for text classification (e.g., [25]). Finally, we have started integrating the classification system into the software used by radiologists for writing the reports. As mentioned before, this will allow the physicians to obtain a "real-time" classification of a report which should then be confirmed or modified. The automatic annotation module will also be used for highlighting the relevant sentences of a report in order to improve visualization. We plan to analyze how/if the automatic annotation and classification of new reports helps the radiologists.

Acknowledgements

The research described in this paper has been partially carried out in the context of the H&W SmartServices project of the University of Brescia. The work of Anne-Lyse Minard has been carried out under research contracts with the University of Brescia and Fondazione Bruno Kessler.

References

[1] Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016;279(2):329–43.
[2] Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. A text processing pipeline to extract recommendations from radiology reports. J Biomed Inform 2013;46(2):354–62.
[3] Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. J Biomed Inform 2013;46(5):869–75.
[4] Khachidze M, Tsintsadze M, Archuadze M. Natural language processing based instrument for classification of free text medical records. BioMed Res Int 2016;2016.
[5] Yim W-w, Kwan SW, Yetisgen M. Classifying tumor event attributes in radiology reports. J Assoc Inf Sci Technol 2017;68(11):2662–74.
[6] Attardi G, Cozza V, Sartiano D. Annotation and extraction of relations from Italian medical records. Proceedings of the 6th Italian information retrieval workshop (IIR 2015) 2015.
[7] Esuli A, Marcheggiani D, Sebastiani F. An enhanced CRFs-based system for information extraction from radiology reports. J Biomed Inform 2013;46(3):425–35.
[8] Marcheggiani D, Sebastiani F. On the effects of low-quality training data on information extraction from clinical reports. J Data Inf Qual 2017;9(1):1:1–1:25. http://dx.doi.org/10.1145/3106235.
[9] Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvist Invest 2007;30(1):3–26. http://dx.doi.org/10.1075/li.30.1.03nad.
[10] McInnes BT, Pedersen T, Carlis J. Using UMLS concept unique identifiers (CUIs) for word sense disambiguation in the biomedical domain. AMIA 2007, American Medical Informatics Association annual symposium, Chicago, IL, USA, November 10–14, 2007 2007.
[11] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn 1997;29(2–3):131–63.
[12] Rokach L, Maimon O. Data mining with decision trees: theory and applications. River Edge, NJ, USA: World Scientific Publishing Co., Inc.; 2008.

[13] Ho TK. Random decision forests. Proceedings of the third international conference on document analysis and recognition (vol. 1) – vol. 1, ICDAR '95. Washington, DC, USA: IEEE Computer Society; 1995. p. 278–82.
[14] Breiman L. Random forests. Mach Learn J 2001;45(1):5–32. http://dx.doi.org/10.1023/A:1010933404324.
[15] Haykin S. Neural networks: a comprehensive foundation. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.; 2007.
[16] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.
[17] Scholkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge, MA, USA: MIT Press; 2001.
[18] Zhou Z-H. Ensemble methods: foundations and algorithms. 1st ed. Chapman & Hall/CRC; 2012.
[19] Qwaider MRH, Minard A-L, Speranza M, Magnini B. Find problems before they find you with AnnotatorPro's monitoring functionalities. Proceedings of the 4th Italian conference on computational linguistics (CLiC-it 2017) 2017.
[20] Girardi C, Bentivogli L, Farajian MA, Federico M. MT-EQuAl: a toolkit for human assessment of machine translation output. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics, Conference System Demonstrations, August 23–29, 2014, Dublin, Ireland 2014:120–3.
[21] Pianta E, Girardi C, Zanoli R. The TextPro tool suite. Proceedings of the sixth international conference on language resources and evaluation (LREC'08), Marrakech, Morocco 2008.
[22] Hall MA, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans Knowl Data Eng 2003;15(6):1437–47. http://dx.doi.org/10.1109/TKDE.2003.1245283.
[23] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl 2009;11(1):10–8.
[24] Baker S, Korhonen A, Pyysalo S. Cancer hallmark text classification using convolutional neural networks. Proceedings of the fifth workshop on building and evaluating resources for biomedical text mining (BioTxtM2016). 2016. p. 1–9. http://www.aclweb.org/anthology/W16-5101.
[25] Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional semantics resources for biomedical text processing. Proceedings of the 5th international symposium on languages, biology and medicine (LBM 2013) 2013:39–44http://lbm2013.biopathway.org/lbm2013proceedings.pdf.