Computational and Structural Biotechnology Journal 24 (2024) 622–629

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

Text mining approach for feature extraction and cartilage disease grade classification using knee MRI radiology reports

Antonio Saverio Valente [c,*,1], Teresa Angela Trunfio [b,1], Marco Aiello [a], Dario Baldi [a], Marilena Baldi [c,1], Silvio Imbò [c], Mario Alessandro Russo [c], Carlo Cavaliere [a,1], Monica Franzese [a,1]

[a] IRCCS SYNLAB SDN, Via E. Gianturco, 113, 80143, Naples, Italy
[b] University of Naples Federico II, Department of Advanced Biomedical Sciences, Via Pansini, 5, 80131, Naples, Italy
[c] GESAN SRL, R&D Department, Via Torino, 14, 81020, San Nicola La Strada, Caserta, Italy

ARTICLE INFO

Keywords:
Structured reporting
Knee cartilage
Magnetic resonance imaging
Neural network
Text mining

ABSTRACT

MRI radiology reporting processes can be improved by exploiting structured and semantically labelled data that can be fed to artificial intelligence (AI) tools. AI-based tools assisting radiology reporting can help to automatically individuate cartilage grading in textual magnetic resonance imaging (MRI) reports, thus supporting clinicians' decisions regarding medical imaging utilisation, diagnosis and treatment. In this study, we extracted information (clinical findings, observations, anatomical regions, etc.) and classified knee cartilage degradation from medical reports utilising transfer-learning techniques applied to the Bidirectional Encoder Representations from Transformers (BERT) model and its variants, pre-trained on an Italian-language corpus. To realise this objective, we used a dataset of 750 MRI knee reports written by three radiologists who contributed to a manual annotation process to perform text classification (TC) and named entity recognition (NER) tasks. The dataset was obtained from an internal database of the IRCCS SYNLAB SDN. Seventy percent of the dataset was used for training, 10% was used for validation and 20% was used for testing. The best-performing configurations for NER and TC tasks were based on the pre-trained BERT model. The macro F1-scores obtained with the NER and TC models are 0.89 and 0.81, respectively. The accuracies calculated on the test set for both tasks are 0.96 and 0.99, respectively.

1. Introduction

Degenerative joint disease of the knee is typically caused by the wear and tear and progressive loss of articular cartilage [33], resulting in severe knee pain that affects the ability to perform normal activities. Damage to the articular knee cartilage can cause pain, inflammation, clicking noise and a catching sensation and can reduce the motility of the joint. Cartilage injuries wider than a centimetre can increase in size over time, involving the subchondral bone and leading to prosthetic surgery. Traditional radiography shows some limitations in the diagnostic stage: it is mainly limited to the bone assessment and articular ratio. The gold standard for assessing knee cartilage injury is MRI, which highlights early signs of knee cartilage degradation involving the subchondral bone [33,1]. Although radiology reports represent an important source of information on the patient's health status, the high operator-related variability and the lack of a standardised writing methodology make automated search for information complex [13]. Several studies have been conducted to improve the standardisation of the information present in reports by introducing *structured reports*. Structured reporting helps define a standard of quality in radiology, leading to greater homogeneity and consistency of reports using a concise lexicon, thus minimising errors, enhancing divisional and departmental branding, improving interdisciplinary communications and encouraging the use of data mining [21] and Natural Language Processing (NLP) [37,28].

In their review, Sloan et al. [34] discuss how the increasing pressure on medical imaging departments, affects radiologists' ability to produce

* Corresponding author.
E-mail addresses: antosaverio.valente@gmail.com (A.S. Valente), teresa.trunfio@gmail.com (T.A. Trunfio), marco.aiello@synlab.it (M. Aiello), dario.baldi@synlab.it (D. Baldi), marilenabaldi95@gmail.com (M. Baldi), silvio.imbo@tabtechnology.it (S. Imbò), mario.russo@gesan.it (M.A. Russo), carlo.cavaliere@synlab.it (C. Cavaliere), monica.franzese@synlab.it (M. Franzese).
[1] These authors contributed equally.

https://doi.org/10.1016/j.csbj.2024.10.003
Received 19 August 2024; Received in revised form 1 October 2024; Accepted 1 October 2024
Available online 5 October 2024

timely and accurate reports; highlighting how AI and Automatic Radiology Report Generation (ARRG) may improve reporting processes. Therefore, applying NLP techniques to unstructured radiological reports can support the analysis and interpretation of medical imaging findings. Some of the main challenges faced by deep learning systems for natural language analysis are related to the individuation of syntactic structures, clearing of ambiguities and identifying the presence of synonyms or possible colloquialisms in the text [31]. This is especially true for the clinical domain, which has a higher degree of complexity with respect to common reading texts, mostly due to case-specific terms and abbreviations [29]. In the musculoskeletal imaging field, several authors have used NLP on radiographic reports for fracture identification. Dai et al. [4] proposed an extraction system called BoneBERT (BERT = Bidirectional Encoder Representations from Transformers) that allows retrieving labelled details of bone fracture from radiology reports written in English, outperforming the conventional rule-based labelling system. Wang et al. [38] validated NLP algorithms on radiology reports (in English), recognising 17 out of 20 fractures. Grundmeier et al. [16] used NLP tools to identify paediatric long bone fractures from radiological reports written in English, reaching accuracy values of up to 95.0%. Jungmann et al. [18] used NLP to automatically analyse the number and distribution of fractures before and during the COVID-19 pandemic by extracting information from major radiographic joint reports written in German. The BERT model performed better than traditional machine learning (ML) approaches on Dutch radiology reports of graded orthopaedic trauma [30]. Relevant results were also obtained for Russian [20] and Chinese radiology reports using the BERT model [24].

As reported in the systematic review of Casey et al. [2], most of the NLP application are available in English due to the abundance of shared resources and tools. In contrast, there are very few Italian-language applications for structured reporting, mainly due to limited data sources and lower scientific interest and dissemination. Esuli et al. [8] exploited two approaches based on conditional random fields to extract information from more than 500 breast radiology reports written in Italian. Galbusera et al. [12] used the BERT model to automatically diagnose spinal disorders from radiographic reports written in Italian, achieving an accuracy of 0.88 - 0.98 and a specificity between 0.84 - 0.99. To date, we are unaware of any work conducted on knee MRI reports written in Italian using transfer-learning techniques applied to BERT, allowing the classification of cartilage degradation. Based on previously reported findings, the aim of this work is twofold: *i*) to create a structured report from free text and *ii*) automatically classify possible osteochondral pathologies with related grading based on radiologist descriptions. Both the extraction of features and disease grading classification from the free-text report will be performed using transfer-learning techniques applied to the BERT model and its variants, pre-trained on an Italian-language corpus.

2. Materials and methods

2.1. Dataset

Medical reports are textual documents containing patient data, examination details and physician observations. This study adheres to the Declaration of Helsinki and was approved by the ethical committee (protocol Big data number 1/20). The dataset was obtained from the IRCCS SYNLAB SDN database, hosted on a Microsoft SQL Server virtual machine. Data entry follows the case report forms, with patient information anonymised. Reports from the last 5 years of clinical activities at the institute were stored in this database in free-text form.

The first step of our research activities has been the creation of the initial dataset considering all the unstructured medical reports related to knee pain stored in the database as sources of information. Several analyses have been performed on the database by the user interface and by implementing dedicated queries to prepare the dataset for the following analysis. We filtered the collected medical reports using specific keyword (e.g. ‘rotula’, ‘menisco’, ‘legamenti crociati’, etc.) to retrieve only those related to knee MRI. To ensure the generalizability of the NLP model, we included reports written by three radiologists, accounting for variations in writing style and maximising the dataset size. A manual check was conducted to ensure that the extracted reports met our criteria. At the end of the extraction and filtering procedure of the dataset, we obtained 250 reports by each of the three radiologists, affording a total of 750 reports related to knee MRI. All these medical reports were then selected and exported in *.csv* format (an example is shown in Table 1).

The dataset was pre-processed to facilitate network learning by removing leading/trailing spaces, special characters (e.g. carriage return), legal/administrative terms, administration of radioactive substances, extra punctuation and replacing dates with a generic pattern (e.g., dd/mm/yyyy). Once this process was completed, reports were saved as text files ready to be used in the annotation phase (see Subsections 2.3.1 and 2.4.1). Finally, after the annotation processes, we created datasets for learning, validation and evaluation of the models and subsequently used them as input for tokenisation and model learning (NER and TC models). To this end, the reports were divided by performing stratification conducted by radiologists, and 70% of the samples were included in the training set, 10% in the validation set and the remaining 20% in the test set.

2.2. BERT and RoBERTa

At the core of BERT is the Transformer architecture, a novel neural network design that introduces the concept of attention. The attention mechanism, introduced by Vaswani et al. [36], was developed to move beyond traditional statistical methods by modelling the relationships between elements within a sequence. This approach allows the model to capture how different parts of the input interact with each other and assess the influence that one part has on another. These interactions that reproduce the input context are represented numerically in an attention matrix [11]. BERT inputs are sentences in which words should be formatted with preliminary operations before being given to the model. In particular, words are chunked (by using WordPiece) and formatted (using a special classification and separation token) to recognise the end of a sentence or the separation between two sentences. Further embedding can be used for specific tasks, such as Next Sentence Prediction (NSP), to help the model distinguish between different segments within the same input sequence, like denoting two separate sentences [15]. The BERT framework is characterised by two steps: *i*) Pre-training and *ii*) Fine-tuning.

RoBERTa (Robustly Optimized BERT Pre-training Approach) is a more robust variant of BERT developed to optimise performance [5]. Specifically, the new model uses dynamic-type masks to improve text encoding and larger batch sizes in the training phase [23]. In addition, the model uses different hyper-parameters and a Byte-Pair Level “tokenization” and is not trained for the NSP task, thus improving results in specific tasks [25].

In this study, we compared the performances of pre-trained and trained BERT and RoBERTa models, by calculating precision, recall, and F1 measure [32] and accuracy [9].

2.3. Named entity recognition

One of the most widely performed tasks for analysing and extracting information from unstructured text is NER. NER identifies and classifies named entities mentioned in a text into pre-defined categories [39]; we see it as a multi-classification problem where each N word in a text is assigned to one of the T-identified categories [10]. In this study, we performed NER using a supervised learning approach to extract information about knee lesions and their characteristics from medical reports. The model used to extract clinical information in a structured format comprises three entity groups according to Sugimoto et al. [35]: *observation*,

Table 1
Samples from medical reports.

Diagnostic Test	Type of test	Report (Italian)	Report (English)
MRI ginocchio destro *(right knee)*	TSE DP su due piani sagittali *(TSE DP on two sagittal planes)*	Regolare morfologia della capsula articolare e delle strutture ligamentose di rinforzo capsulare. Fibrocartilagini meniscali di normale morfologia e segnale. Apprezzabili da inserzione ad inserzione i legamenti crociati esenti da alterazioni del segnale. Regolare morfologia e segnale delle strutture ossee articolari con fisiologica rappresentazione del rivestimento condrale femoro-tibiale e femoro-rotuleo. Minimo versamento intraarticolare.	Normal morphology of the articular capsule and the capsular reinforcing ligamentous structures. Meniscal fibrocartilages show normal morphology and signal. The cruciate ligaments are appreciable from insertion to insertion and are free from signal alterations. Normal morphology and signal of the articular bone structures with a physiological representation of the femorotibial and femoropatellar cartilage covering. Minimal intra-articular effusion.
MRI ginocchio destro *(right knee)*	TSE DP su due piani sagittali e TSE T2 su piani assiali *(TSE DP on two sagittal planes and TSE T2 on axial planes)*	Reperto RM limitato dalla presenza di protesi bicompartimentale che inficia la diagnosi. Con tali limiti si segnala Tendinosi del tendine del quadricipite e del tendine rotuleo. Ispessimento sinoviale anteriore. Usura condrale femoro-rotulea di III grado con lacune di riassorbimento. Discreto versamento. Legamenti crociati e menischi non valutabili.	MRI findings are limited by the presence of a bicompartmental prosthesis, which impairs the diagnosis. Within these limitations, quadriceps tendon and patellar tendon tendinosis is noted. Anterior synovial thickening. Grade III femoropatellar cartilage wear with resorption lacunae. Moderate effusion. Cruciate ligaments and menisci are not assessable.
MRI ginocchio destro *(right knee)*	TSE DP su due piani sagittali e TSE T2 su piani assiali *(TSE DP on two sagittal planes and TSE T2 on axial planes)*	Esiti di frattura dell'apofisi tibiale anteriore consolidata in vizio parziale con tendinosi del rotuleo in sede inserzionale Fibrocartilagini meniscali di normale morfologia e segnale. Apprezzabili da inserzione ad inserzione i legamenti crociati esenti da alterazioni del segnale. Lieve versamento intraarticolare. Moderato ispessimento sinoviale anteriore. Fibroma al terzo distale di femore e terzo prossimale di tibia di circa 25 mm di diametro craniocaudale In via collaterale si segnalano fibromi al terzo distale di femore e tibia sinistra ripettivamente di 26 e 22 mm.	Results of a fracture of the anterior tibial tuberosity, consolidated with partial malunion, with tendinosis of the patellar tendon at the insertion site. Meniscal fibrocartilages show normal morphology and signal. The cruciate ligaments are appreciable from insertion to insertion and are free from signal alterations. Mild intra-articular effusion. Moderate anterior synovial thickening. Fibroma in the distal third of the femur and proximal third of the tibia, approximately 25 mm in craniocaudal diameter. Collaterally, fibromas are noted in the distal third of the left femur and tibia, measuring 26 and 22 mm, respectively.

clinical finding and *modifier entities*. Starting from an observation, characterised by different "modifier entities" (i.e. size, change and anatomic location entities), a radiologist can define the clinical findings.

2.3.1. Annotation process

The text was annotated manually by a clinical expert of the IRCCS SYNLAB SDN, associating the words of each medical report with the corresponding tag. To achieve proper NER, we defined all the keywords and their possible combinations, which were then used as references during the annotation and training phase. Pre-processed data (Section 2.1) has been annotated to identify the relationships between the words in the input medical reports and the specified entities. To better disclose the information model for the NER task, we first defined the entities to extract the above-mentioned *observations*, *clinical findings* and *modifiers* and the tags to use. In particular, we have identified the following *tags*:

- OBS (Observation). Represents a clinical observation and refers to measuring, questioning, evaluating or otherwise observing a patient or a specimen. Observations are typically characterised by terms describing specific test results, allowing clinicians to formulate a diagnosis.
- CLI (Clinical Finding). Represents a specific pathological condition, typically defined by a particular name with no further terms characterising it other than related to certainty, extent and anatomical location.
- LOC (Anatomical Location). Represents an adjective or noun modifier of anatomical location and is used to denote the area where an observation is made or a disease is found.
- CER (Certainty). Represents a certainty modifier that determines whether an entity is absent.
- CHG (Change). Represents a modifier of change in the status of an entity typically in the form of an adjective.
- CHS (Characteristics). Represents a feature modifier typically characterising observation entities by adjectives.
- SIZ (Size). Is a size modifier and mostly identifies numerical values and reported units of measurement.

Fig. 1 shows an example of structured information extraction from a medical report in our dataset.

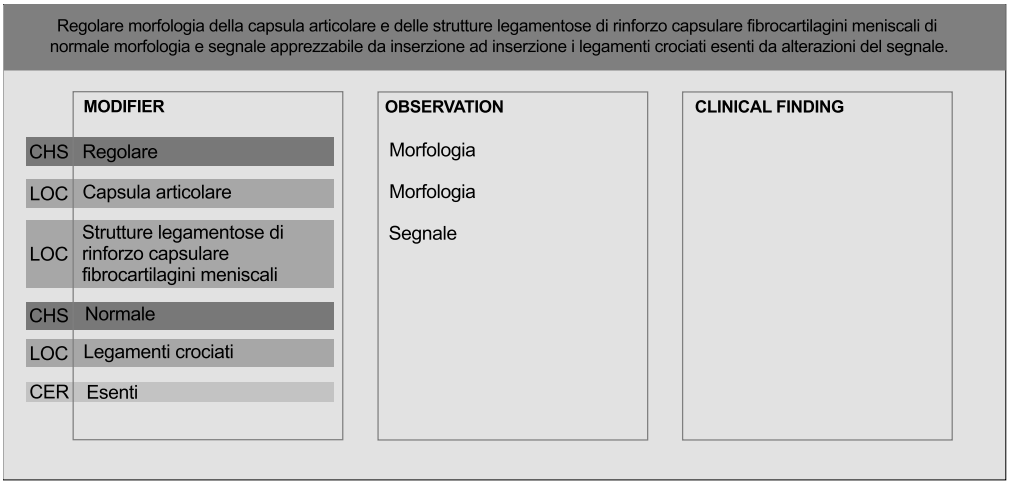
2.4. TC

The second goal of this study has been the automatic classification of cartilage injury by associating each medical report with a severity class. To this aim, we have developed a model to solve a TC task by exploiting the same dataset described in Section 2.1. TC is an NLP technique that assigns a classification label or category to pieces of unstructured text. TC solves a multiclass classification problem by taking a tokenised sentence as input and assigning a category to it. In this study, we solved the classification problem of medical reports by using a model trained with a supervised learning approach. Another annotation process has been necessary to create a dataset containing pairs of sentences\grade (read: medical report\severity).

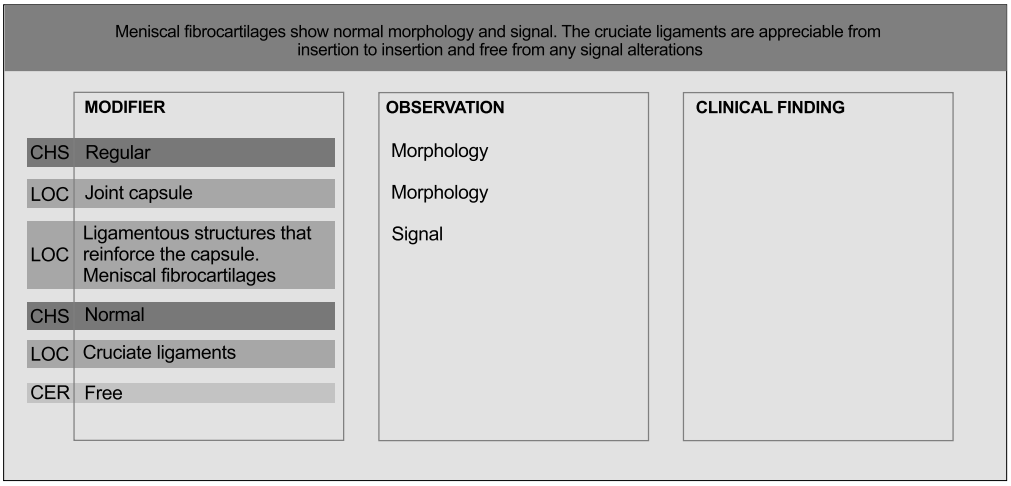
2.4.1. Annotation process

The manual annotation phase for TC has consisted of labelling the medical reports with different severity of cartilage impairment, including the class indicating the disease absence. The following annotation classes are reported:

- I. Chondral softening;
- II. Second Grade superficial lesions extending down to $< 50\%$ of cartilage depth;
- III. Third grade cartilage defects extending down to 50% of depth but not through subchondral bone;



(a)



(b)

Fig. 1. Example of report annotation in a structured format: in Italian (*a*); and its corresponding interpretation in English (*b*).

Table 2
Training intervals and optimal values identified for hyper-parameters in NER and TC tasks.

Training Parameters			
Hyper-parameter	Intervals	Optimal Value	
		[NER]	[TC]
epochs	5, 8, 13	13	13
batch-size	4, 8, 16	4	8
initial learning rate	1e-3, 1e-4, 1e-5	1e-5	1e-5

- IV. Fourth grade ulceration through subchondral bone;
- NC. For sentences not referring to cartilage assessment.

The annotated dataset was divided into 70% for the training set, 10% for the validation set and 10% for the testing set.

2.5. Fine-tuning of language models

As shown in Table 2, for each hyper-parameter, we report the intervals of the search area and the optimal values identified for both NER and TC tasks.

To mitigate the effects of the limited availability of labelled data for the TC task, we fine-tuned the hyper-parameters during the training phase, focusing on class balancing and adjusting specific parameters of the *BertForSequenceClassification* model, such as attention probabilities and dropout rates (*probs_dropout_prob* and *dropout_prob*). Starting with a default value of 0.1, we incrementally adjusted both parameters by 0.1 steps, ultimately setting them to the optimal value of 0.4. During the training phase, we exploited the following configurations:

1. $n = 64$ is the size of the tokenised sentences used as input to the models;
2. Adam optimisation algorithm for the training functions;
3. Gradient clipping technique with a threshold value of 10 to prevent exploding gradient in the training functions.

2.6. Software pipeline

In this section, we describe the computation pipeline to process the medical report, shown in the flow diagram in Fig. 2. The first stage of the pipeline, common to both tasks performed in this study, has been indicated as data acquisition in Fig. 2. The data acquisition stage included all the preliminary activities required to get the information of interest from the SDN database. In detail, medical reports from the SDN database were filtered to obtain the subset of interest, then extracted, cleaned and stratified (as described in Section 2.1) considering the different writing styles of radiologists. At this point, we performed the annotation process, which was done manually for the two tasks under the supervision of an experienced radiologist. The output of the data acquisition and

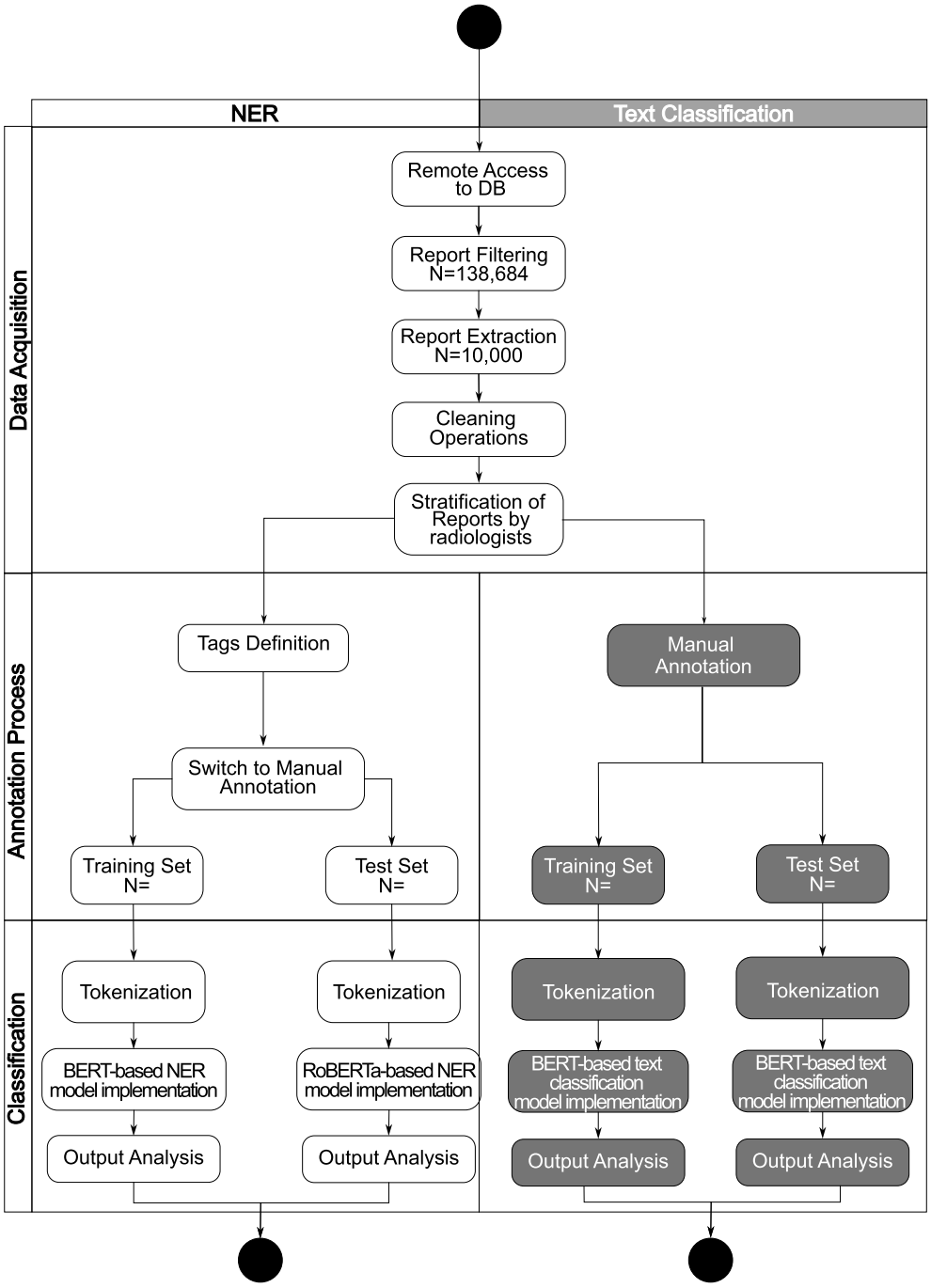


Fig. 2. Software pipeline implemented for the execution of NER and TC tasks.

annotation process stage allowed the creation of the dataset serving as input to the models.

The final stage, called classification, was characterised mainly by the implementation of neural networks. Although both the tasks are based on the same BERT model, different final neural network layers have been developed to handle each stage (Fig. 2). In fact, for NER, the classification has been performed by identifying individual words as inputs. For TC, the inputs consist of whole sentences. Therefore, before implementing the models, different tokenisation operations were performed for the BERT-based NER model, BERT-based TC model, RoBERTa-based NER model and RoBERTa-based TC model. The pipeline ends with the training, validation and evaluation of the neural networks.

The proposed AI solution was implemented using the programming language *Python*. During the development phase, we exploited *PyCharm*, an IDE with an interactive console integrated with the IPython Notebook supporting *Anaconda*, a system used for managing virtual environments. At the following we have uploaded the models used during the experimental validation. At the following *GitHub* link (*https://github.com/marilena-baldi/text-mining-mri*), you can find the AI models that have been developed for the TC and NER tasks.

3. Results

An important component in the design of neural networks is the determination of their structure and parameters. However, these elements depend on the particular problem, and their value considerably influences performance. Typically, these parameters are chosen using heuristic rules or are manually tuned [7].

Table 3
Experimental results for NER and TC. Metrics are macro averaged.

Task	Model	Language Model	Precision	Recall	F1 score
NER	BERT	pre-trained	0.91	0.88	0.89
	BERT	trained	0.84	0.89	0.86
	RoBERTa	trained	0.81	0.82	0.81
TC	BERT	pre-trained	0.86	0.81	0.82
	BERT	trained	0.83	0.80	0.81
	RoBERTa	trained	0.83	0.64	0.60

Table 4
Experimental results for single categories and the classification stage.

Task	Class	Precision	Recall	F1 score	Accuracy
NER	CER	0.93	0.98	0.95	0.96
	CHG	0.95	0.58	0.72	
	CHS	0.89	0.90	0.89	
	CLI	0.90	0.94	0.92	
	LOC	0.92	0.92	0.92	
	OBS	0.91	0.91	0.91	
	SIZ	0.89	0.95	0.92	
	micro average	0.91	0.91	0.91	
	macro average	0.91	0.88	0.89	
TC	I	0.75	0.80	0.77	0.81
	II	0.76	0.79	0.78	
	III	0.80	0.80	0.80	
	IV	0.81	0.79	0.80	
	NC	0.90	0.86	0.88	
	micro average	0.81	0.81	0.81	
	macro average	0.80	0.81	0.81	

Different configurations have been analysed to solve the two tasks (NER and TC) by varying the implemented network model or using different types of models (pre-trained and trained). For both tasks, different models based on BERT and RoBERTa have been implemented. BERT has been used for transfer-learning utilising a pre-trained language model on an Italian-language corpus and a model directly trained on the corpus of medical reports relevant to this study. RoBERTa has been directly trained on the corpus of medical reports relevant to this study. To reduce any bias (depending on the order of the training samples) and speed up the convergence of the algorithms, the training dataset was subjected to shuffling at each iteration. The validation datasets have been used to search the optimal hyper-parameters for both tasks with respect to each model (models pre-trained\trained from scratch based on BERT and models trained from scratch based on RoBERTa). For both tasks, the optimal model was based on pre-trained BERT.

Table 3 shows the results obtained for NER and TC, considering the best hyper-parameter configurations. We evaluated the performance of NER and TC models by calculating precision, recall, F1 score and accuracy. Our experiments show that the best-performing configurations were those based on the pre-trained BERT model. The macro F1-scores obtained with the NER and TC models were 0.89 and 0.82, respectively.

In particular, considering the configuration that achieved the best performance (i.e. BERT pre-trained on an Italian corpus), the results for the two tasks are given in Table 4. The accuracy calculated on the test set for NER was 0.96, whereas it was 0.81 for TC.

Despite the challenges posed by the clinical domain's specific semantic structures and terminologies, our system achieved good results. This highlights the effectiveness of pre-training on large-scale corpora in capturing relevant contextual information for accurate classification tasks.

4. Discussion

MRI is the gold standard for knee joint assessment, detecting early cartilage degradation and subchondral bone involvement. Furthermore, coupling the MRI information with the clinical outcomes in large databases can promote the development of ML models, which can help support the diagnosis. However, subjective variability in radiological reports and lack of standardisation hinder automated information extraction of clinical outcomes. Structured reporting using AI and NLP tools aims to improve consistency and data mining, aiding clinical decision making. To date, all recent state-of-the-art models in NLP appear to rely on Transformer-based architectures [14]. The use of Transformer allowed addressing the NLP problems without the use of recurrent neural networks (RNNs), thus achieving benefits in terms of computational parallelism, time efficiency and model robustness [19]. One of the most commonly used frameworks for generating language models for NLP tasks is BERT [6]. In contrast to directional models, which sequentially read the text input (left-to-right or right-to-left), the Transformer encoder at the base of the BERT architecture allows reading the entire sequence of words at once, indicating its bidirectional nature. The advantage of BERT is its ease of use, which involves adding just one output layer to the existing neural architecture to obtain text models that surpass the inaccuracy of all existing ones (like Word2Vec and Glove) on several natural text processing problems [22]. Furthermore, López-Úbeda et al. [27] compared the effectiveness of RoBERTa, convolutional neural networks (CNNs) and ChatGPT in detecting unexpected findings in radiology reports. They reported that RoBERTa achieved the highest accuracy and F1 score, outperforming CNN and ChatGPT in identifying critical, unexpected findings from the radiology text. In another study, López-Úbeda et al. [26] addressed the critical problem of accurate summarisation in radiology reports by comparing various large language model–based approaches for automatic summary generation. They employed two language models—Text-to-Text Transfer Transformer (T5) and Bidirectional and Auto-Regressive Transformers (BART)—and compared them with an Recurrent Neural Network (RNN) on a dataset of 15,508 retrospective knee MRIs. Summaries produced by the T5 model were similar to those produced by a radiologist, with approximately 70% similarity in fluency and consistency.

This study used NER and TC techniques to analyse knee medical reports, extracting keywords and classifying cartilage severity using ML models to propose a general method to yield labelled datasets from clinical radiological repositories. BERT and RoBERTa configurations were tested, with BERT (pre-trained on an Italian corpus) affording the best performance.

Optimal hyper-parameters were identified for epochs, batch sizes and learning rates. The pre-trained BERT model achieved good performance: for NER, a macro F1-score of 0.89 (precision = 0.91 and recall = 0.88); for TC, a macro F1-score of 0.82 (precision = 0.86 and recall = 0.81).

Results were further divided into specific categories (CER, CHG, CHS, CLI, LOC, OBS and SIZ), with the NER task reporting an accuracy of 0.96 on the test set. To characterise the cartilage degradation staging, for the TC task, different classes (I, II, III, IV and NC) were evaluated, with an overall accuracy of 0.81 on the test set.

To the best of our knowledge, few studies in the literature have focused on applying NLP to knee clinical reports. In knee imaging, Chen et al. [3] presented promising results of a BERT model trained on radiological reports to identify cartilage lesions in patients with osteoarthritis (OA). Hassanpour et al. [17] evaluated the performance of an NLP model on two testing datasets, classifying English-language knee MRI reports with excellent accuracy (with an F1 score of >77.0%) even on an inde-

pendent untrained dataset. Our work presented a substantially different approach and different outcomes with respect to previous studies aiming to specifically apply NLP in classifying knee radiology reports [3,17]. Chen et al. [3] used a pre-trained BERT model fine-tuned on knee radiology reports, while Hassanpour et al. [17] developed an SVM model for free-text knee MRI reports. All studies aimed to identify cartilage lesions in patients with OA. In our study, we classified lesion severity into five classes. Chen et al. [3] employed the WORMS system for binary classification, whereas Hassanpour et al. [17] used a manual classification scheme. Overall, all these studies achieved comparable performance: our model achieved a 0.81 accuracy, [3] achieved a 0.89 accuracy and [17] achieved a 0.84 accuracy.

Our study demonstrated the suitability of pre-trained models, especially BERT, in processing radiological reports. Despite limited labelled data, these models performed well, showcasing the potential of transfer learning for medical TC and entity recognition. Results highlight the importance of large-scale pre-trained models in capturing contextual information for accurate medical tasks.

Additionally, we achieved satisfactory results in structuring medical reports and identifying knee cartilage disease severity, suggesting that even with data scarcity of labelled data, state-of-the-art models such as BERT can excel with minimal expert labelling. Although our findings are promising, some limitations are present and further investigations are needed. Firstly, the dataset is relatively small, which may limit the generalizability of the findings to larger or datasets. Additionally, all reports were written by only three radiologists, which could introduce a bias based on individual reporting styles. A larger number of annotators would likely improve the robustness of the model and reduce any potential variability in the manual annotation process. Expanding the dataset and incorporating domain-specific knowledge could improve the performance of the models and enable better recognition of complex medical conditions.

5. Conclusion

We explored the application of the NER and TC techniques in the analysis of medical reports related to knee osteoarticular pathologies to extract keywords and classify the corresponding severity degrees. Our study demonstrates that the combination of advanced models and proper training techniques allows accurate information extraction for structuring reports and identifying clinical conditions.

CRediT authorship contribution statement

Antonio Saverio Valente: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Teresa Angela Trunfio:** Writing – original draft, Methodology, Investigation, Conceptualization. **Marco Aiello:** Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Dario Baldi:** Methodology, Investigation, Formal analysis, Conceptualization. **Marilena Baldi:** Validation, Software, Data curation, Conceptualization, Formal analysis, Investigation, Methodology. **Silvio Imbò:** Validation, Software, Data curation. **Mario Alessandro Russo:** Validation. **Carlo Cavaliere:** Writing – original draft, Supervision, Investigation, Formal analysis, Data curation, Conceptualization. **Monica Franzese:** Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This paper has been supported by a Ricerca Corrente project from the Italian Ministry of Health. The activities described in this paper have also been partially carried out under the RIGOLETTO project, “Realizzazione pIattaforma GestiOne inteLligEnTe pazienTe Oncologico”, funded by the Ministry of Enterprises and Made in Italy, Project No.: F/310339/01/X5.

References

[1] Bellelli A, Silvestri E, Barile A, Albano D, Aliprandi A, Caudana R, et al. Position paper on magnetic resonance imaging protocols in the musculoskeletal system (excluding the spine) by the Italian college of musculoskeletal radiology. Radiol Med 2019;124:522–38. https://doi.org/10.1007/s11547-019-00992-3.

[2] Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. BMC Med Inform Decis Mak 2021;21:1–18. https://doi.org/10.1186/s12911-021-01533-7.

[3] Chen L, Shah R, Link T, Bucknor M, Majumdar S, Pedoia V. Bert model fine-tuning for text classification in knee oa radiology reports. Osteoarthr Cartil 2020;28:S315–6. https://doi.org/10.1016/j.joca.2020.02.488.

[4] Dai Z, Li Z, Han L. Bonebert: a bert-based automated information extraction system of radiology reports for bone fracture detection and diagnosis. In: International symposium on intelligent data analysis. Springer; 2021. p. 263–74. doi.

[5] Delobelle P, Winters T, Berendt B. Robbert: a Dutch Roberta-based language model. arXiv preprint. arXiv:2001.06286. doi: https://doi.org/10.48550/arXiv.2001.06286, 2020.

[6] Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. arXiv:1810.04805. doi: https://doi.org/10.48550/arXiv.1810.04805, 2018.

[7] Diaz GI, Fokoue-Nkoutche A, Nannicini G, Samulowitz H. An effective algorithm for hyperparameter optimization of neural networks. IBM J Res Dev 2017;61:9:1–911:. https://doi.org/10.1147/JRD.2017.2709578.

[8] Esuli A, Marcheggiani D, Sebastiani F. An enhanced crfs-based system for information extraction from radiology reports. J Biomed Inform 2013;46:425–35. https://doi.org/10.1016/j.jbi.2013.01.006. doi.

[9] Fawcett T. An introduction to roc analysis. Pattern Recognit Lett 2006;27:861–74. https://doi.org/10.1016/j.patrec.2005.10.010. doi.

[10] Galar M, Fernández A, Barrenechea E, Bustince H, Herrera F. An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and one-vs-all schemes. Pattern Recognit 2011;44:1761–76. https://doi.org/10.1016/j.patcog.2011.01.017. https://www.sciencedirect.com/science/article/pii/S0031320311000458.

[11] Galassi A, Lippi M, Torroni P. Attention in natural language processing. IEEE Trans Neural Netw Learn Syst 2020;32:4291–308. https://doi.org/10.1109/TNNLS.2020.3019893.

[12] Galbusera F, Cina A, Bassani T, Panico M, Sconfienza LM. Automatic diagnosis of spinal disorders on radiographic images: leveraging existing unstructured datasets with natural language processing. Glob Spine J 2021. https://doi.org/10.1177/21925682211026910.

[13] Ganeshan D, Duong PAT, Probyn L, Lenchik L, McArthur TA, Retrouvey M, et al. Structured reporting in radiology. Acad Radiol 2018;25:66–73. https://doi.org/10.1016/j.acra.2017.08.005.

[14] Gillioz A, Casas J, Mugellini E, Khaled OA. Overview of the transformer-based models for nlp tasks. In: 2020 15th conference on computer science and information systems (FedCSIS); 2020. p. 179–83.

[15] van der Goot R, Müller-Eberstein M, Plank B. Frustratingly easy performance improvements for low-resource setups: a tale on BERT and segment embeddings. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, et al., editors. Proceedings of the thirteenth language resources and evaluation conference. Marseille, France: European Language Resources Association; 2022. p. 1418–27. https://aclanthology.org/2022.lrec-1.152.

[16] Grundmeier RW, Masino AJ, Casper TC, Dean JM, Bell J, Enriquez R, et al. Identification of long bone fractures in radiology reports using natural language processing to support healthcare quality improvement. Appl Clin Inform 2016;7:1051–68. https://doi.org/10.4338/aci-2016-08-ra-0129.

[17] Hassanpour S, Langlotz CP, Amrhein TJ, Befera NT, Lungren MP. Performance of a machine learning classifier of knee mri reports in two large academic radiology practices: a tool to estimate diagnostic yield. Am J Roentgenol 2017;208:750–3. https://doi.org/10.2214/AJR.16.16128.

[18] Jungmann F, Kämpgen B, Hahn F, Wagner D, Mildenberger P, Düber C, et al. Natural language processing of radiology reports to investigate the effects of the covid-19 pandemic on the incidence and age distribution of fractures. Skelet Radiol 2022;51:375–80. https://doi.org/10.1007/s00256-021-03760-5.

[19] Karita S, Chen N, Hayashi T, Hori T, Inaguma H, Jiang Z, et al. A comparative study on transformer vs rnn in speech applications. In: 2019 IEEE automatic speech recognition and understanding workshop (ASRU). IEEE; 2019. p. 449–56.

[20] Kivotova E, Maksudov B, Kuleev R, Ibragimov B. Extracting clinical information from chest X-ray reports: a case study for Russian language. In: 2020 international conference nonlinearity, information and robotics (NIR). IEEE; 2020. p. 1–6.
[21] Kohli A, Castillo S, Thakur U, Chhabra A. Structured reporting in musculoskeletal radiology. In: Seminars in musculoskeletal radiology. Thieme Medical Publishers, Inc.; 2021. p. 641–5.
[22] Koroteev M. Bert: a review of applications in natural language processing and understanding. arXiv preprint. arXiv:2103.11943. doi: https://doi.org/10.48550/arXiv.2103.11943, 2021.
[23] Liao W, Zeng B, Yin X, Wei P. An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta. Appl Intell 2021;51:3522–33. https://doi.org/10.1007/s10489-020-01964-1. doi.
[24] Liu H, Zhang Z, Xu Y, Wang N, Huang Y, Yang Z, et al. Use of bert (bidirectional encoder representations from transformers)-based deep learning method for extracting evidences in Chinese radiology reports: development of a computer-aided liver cancer diagnosis framework. J Med Internet Res 2021;23:e19689. https://doi.org/10.2196/19689.
[25] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: a robustly optimized bert pretraining approach. arXiv preprint. arXiv:1907.11692. doi: https://doi.org/10.48550/arXiv.1907.11692, 2019.
[26] López-Úbeda P, Martín-Noguerol T, Díaz-Angulo C, Luna A. Evaluation of large language models performance against humans for summarizing mri knee radiology reports: a feasibility study. Int J Med Inform 2024;187:105443. https://doi.org/10.1016/j.ijmedinf.2024.105443. https://www.sciencedirect.com/science/article/pii/S1386505624001060.
[27] López-Úbeda P, Martín-Noguerol T, Escartín J, Luna A. Role of natural language processing in automatic detection of unexpected findings in radiology reports: a comparative study of roberta, cnn, and chatgpt. Acad Radiol 2024. https://doi.org/10.1016/j.acra.2024.07.057. https://www.sciencedirect.com/science/article/pii/S1076633224005622.
[28] Mallio CA, Sertorio AC, Bernetti C, Beomonte Zobel B. Large language models for structured reporting in radiology: performance of gpt-4, chatgpt-3.5, perplexity and bing. Rradiol Med 2023;128:808–12. https://doi.org/10.1007/s11547-023-01651-4.
[29] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011;18:544–51. https://doi.org/10.1136/amiajnl-2011-000464.
[30] Olthof AW, Shouche P, Fennema E, IJpma F, Koolstra R, Stirler V, et al. Machine learning based natural language processing of radiology reports in orthopaedic trauma. Comput Methods Programs Biomed 2021;208:106304. https://doi.org/10.1016/j.cmpb.2021.106304.
[31] Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learn Syst 2021;32:604–24. https://doi.org/10.1109/TNNLS.2020.2979670.
[32] Powers DMW. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. URL: https://api.semanticscholar.org/CorpusID:3770261. https://doi.org/10.48550/arXiv.2010.16061, 2011.
[33] Sharma L. Osteoarthritis of the knee. N Engl J Med 2021;384:51–9. https://doi.org/10.1056/NEJMcp1903768. pMID: 33406330.
[34] Sloan P, Clatworthy P, Simpson E, Mirmehdi M. Automated radiology report generation: a review of recent advances. IEEE Rev Biomed Eng 2024. https://doi.org/10.1109/RBME.2024.3408456.
[35] Sugimoto K, Takeda T, Oh JH, Wada S, Konishi S, Yamahata A, et al. Extracting clinical terms from radiology reports with deep learning. J Biomed Inform 2021;116:103729. https://doi.org/10.1016/j.jbi.2021.103729. https://www.sciencedirect.com/science/article/pii/S1532046421000587.
[36] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30. https://doi.org/10.48550/arXiv.1706.03762.
[37] Vosshenrich J, Nesic I, Cyriac J, Boll DT, Merkle EM, Heye T. Revealing the most common reporting errors through data mining of the report proofreading process. Eur Radiol 2021;31:2115–25. https://doi.org/10.1007/s00330-020-07306-6.
[38] Wang Y, Mehrabi S, Sohn S, Atkinson EJ, Amin S, Liu H. Natural language processing of radiology reports for identification of skeletal site-specific fractures. BMC Med Inform Decis Mak 2019;19:23–9. https://doi.org/10.1186/s12911-019-0780-5.
[39] Zitouni I. Natural language processing of semitic languages. Springer Berlin, Heidelberg; 2014.