Clinical Text Classification in Cancer Real-World Data in Spanish

Francisco J. Moreno-Barea[1(✉)], Héctor Mesa[1], Nuria Ribelles[2], Emilio Alba[2], and José M. Jerez[1]

[1] Departamento de Lenguajes y Ciencias de la Computación, Escuela Técnica Superior de Ingeniería Informática, Universidad de Málaga, Málaga, Spain
fjmoreno@lcc.uma.es

[2] Unidad de Gestión Clínica Intercentros de Oncología, Hospitales Universitarios Regional y Virgen de la Victoria, Málaga, Spain

Abstract. Healthcare systems currently store a large amount of clinical data, mostly unstructured textual information, such as electronic health records (EHRs). Manually extracting valuable information from these documents is costly for healthcare professionals. For example, when a patient first arrives at an oncology clinical analysis unit, clinical staff must extract information about the type of neoplasm in order to assign the appropriate clinical specialist. Automating this task is equivalent to text classification in natural language processing (NLP). In this study, we have attempted to extract the neoplasm type by processing Spanish clinical documents. A private corpus of 23,704 real clinical cases has been processed to extract the three most common types of neoplasms in the Spanish territory: breast, lung and colorectal neoplasms. We have developed methodologies based on state-of-the-art text classification task, strategies based on machine learning and bag-of-words, based on embedding models in a supervised task, and based on bidirectional recurrent neural networks with convolutional layers (C-BiRNN). The results obtained show that the application of NLP methods is extremely helpful in performing the task of neoplasm type extraction. In particular, the 2-BiGRU model with convolutional layer and pre-trained fastText embedding obtained the best performance, with a macro-average, more representative than the micro-average due to the unbalanced data, of 0.981 for precision, 0.984 for recall and 0.982 for F1-score.

Keywords: Text Classification · Natural Language Processing · Electronic Health Records · Neoplasm cancer · Spanish

1 Introduction

Public healthcare systems face numerous challenges, including their sustainability, variability in healthcare practice and the need to improve the patient experience, among others. Evidence-based medicine is based on clinical research and

I. Rojas et al. (Eds.): IWBBIO 2023, LNBI 13919, pp. 482–496, 2023.
https://doi.org/10.1007/978-3-031-34953-9_38

its main tool, randomized clinical trials (RCTs). However, nowadays health outcomes research also includes the collection, compilation and analysis of data generated outside RCTs, in what is known as real-world data (RWD), which in recent years has acquired a growing and renewed interest beyond the classic observational, naturalistic or pragmatic studies, which suffer from significant biases. The widespread, systematic, exhaustive, high quality and transparent collection of data by clinicians in electronic health records, whether these are conventional databases or, more commonly, electronic health records (EHR). Their transformation into useful information of value to the clinician provides a body of knowledge known as Real-World Evidence.

However, the gradual adoption of the EHRs as a key component of healthcare systems raises a number of issues, some of which remain unresolved. EHRs store information of a heterogeneous nature in a variety of formats, including open text documents, such as clinical notes or radiology reports, that contain information related to diagnoses, treatments, or clinical procedures [27]. However, the unstructured nature of these open text fields makes the task of automatically extracting relevant concepts from them particularly difficult, and manual concept extraction is non-reusable, time-consuming and costly [18].

Focusing on a specific medical area, a recurrent problem in oncology clinical analysis units regarding the preparation of the first visit report is the lack of time of the clinical staff to complete the information in the structured fields corresponding to the type of neoplasm, location, histology, etc. This makes subsequent access to the information and exploitation of the results extremely difficult. In other words, the information exists, but it is in text format (not structured) within the EHR information and is not stored in a specific electronic field. The automatic neoplasm type extraction of the text corresponding to the patient's EHR is a key task, allowing the oncology analysis unit to immediately refer the patient to the appropriate specialist.

This process eventually becomes a text classification task. Text classification is a classic problem in natural language processing (NLP). This task is defined as the assignment of text units to one or more categories according to the content and semantics present in the text. These text units can be sentences, questions, paragraphs and, as will be addressed in this study, documents. Text classification is commonly used in marketing, human resources and social analysis tasks such as sentiment analysis (products, companies, online and social media) or news categorisation. Text classification has also proved useful in natural language understanding tasks such as question answering (QA).

Due to the attention this task has received and the increasing amount of textual data, NLP techniques have been applied to the automatic classification of free-text clinical reports in recent years. Approaches to text classification can be divided into three categories: rule-based methods, machine learning (ML) methods and deep learning (DL) methods. The rule-based systems for clinical text classification rely on a large number of manually constructed patterns or rules [15,19]. However, since rule-based methods are not reproducible, studies have focused on ML algorithms for this task. ML methods used include decision tree (DT), naive bayes (NB), support vector machines (SVM) and random forest (RF)

[6,7,10,26]. Finally, with the increased ability to collect large data sets, DL-based methods have become the state-of-the-art (SOTA) for various NPL tasks. Architectures based on convolutional and recurrent neural networks and transformers show impressive results in the text classification task. Models such as long short-term memory (LSMT) [8] and gated-recurrent-units (GRU) [5], including variations such as Bidirectional-LSTM (Bi-LSTM) [12] or Convolutional-LSTM (C-LSTM) [23], and large pre-trained language models with layers of multi-head self-attention architectures [28], have been applied to numerous clinical text classification tasks [2,11,14,29,33].

However, most of the existing studies in the specific literature refer only to texts in English, due to the scarce availability of linguistic corpora annotated with clinical coding information in other languages. Since Spanish is the second most spoken language in the world in terms of number of native speakers [30], there is a need to apply medical NLP methodologies focused on this language. For this text classification task, we have access to the Galén system [22,27], a repository of 60,000 real-world clinical EHRs. The use of this clinical linguistic corpus allows us to obtain reliable information on frequently used words in oncology, as well as grammatical and contextual information in this specific field. Furthermore, the availability of the neoplasm annotations in Gálen for supervised learning allows us to serve as an artificial intelligence laboratory on cancer for the development of NLP models, deploying them in national or international hospitals in Spanish where the neoplasm annotations are not available.

Considering all the above aspects, in this work we propose to advance in the application of NLP models for the automatic extraction of neoplasm type from EHR written in Spanish. The classification algorithms assign to each document the probability of belonging to one of the three most common neoplasms in the Galén information system, as a representation of the Spanish region of Málaga: breast, colorectal and lung; or to another type of neoplasm. ML and DL models studied herein represent SOTA in text classification tasks [13,32], such as RNNs used in conjunction with CNN and embedding models. However, to the best of our knowledge, this is the first study that examines the application of NLP models to the problem of extracting information about the neoplasm suffered by a patient using real-world medical texts in Spanish.

2 Materials

This section describes the corpus used to perform the text classification task. The automatic classification of clinical texts requires a prior manual analysis of the documents for their collection and correct labelling. In this sense, the research team was able to obtain quality-assured information in a simpler way thanks to the availability of Galén system [22,27], an integrated software system in oncology centres in the province of Málaga, Spain. The Galén system collects the EHRs of more than 60,000 oncology patients from the *Hospital Regional Universitario* and the *Hospital Universitario Virgen de la Victoria* in Málaga, Spain, with information completed both in real time and by dedicated staff.

A corpus of EHRs containing an associated neoplasm and containing more than 500 words was selected from the information available in the database, for a total of 23, 704 documents. Each document includes the demographic information, first visit and all information from the remaining episodes (consultation, emergency visit or comments).

After selecting the corpus for the text classification task, the neoplasm labels were processed to group them into breast, lung, colorectal and other neoplasms. The category "other" includes documents on head/neck, liver, prostate, uterus, non-Hodgkin's lymphoma, thyroid, stomach/esophagus and other neoplasms. The selected documents were tokenized, making several decisions to reduce the size of the vocabulary and maximise the inference of contextual relationships. The tokenization was case-insensitive and easily recognisable expressions were replaced by special tokens. In addition, authorised experts obfuscated the documents to maintain anonymity of the real-world EHR for processing. The obfuscation was a bijective transformation of the characters with the additional aim of not losing the properties of n-grams in embedding models.

Table 1 shows the distribution of the neoplasms present in the selected corpus. For the different training, validation and test sets, the columns show the absolute number (abs) of documents for each neoplasms considered and their relative frequency (rel). The majority of neoplasms in the Galén corpus represent the category other (41.9%), although they are not individually sufficiently representative. The most common neoplasm in the corpus is breast cancer (27.4%), while lung and colorectal neoplasms are in the minority but well represented in relation to the rest. In addition, an almost perfect stratification is observed in the training/validation/testing division for the correct evaluation of NLP experimental results.

Table 1. Number and percentage of documents per neoplasm in each corpus subset: training, validation and test.

Neoplasm	Train		Validation		Test	
	abs	rel	abs	rel	abs	rel
Breast	5266	.2743	576	.2699	649	.2738
Lung	2731	.1422	302	.1415	338	.1426
Colorectal	3149	.1640	354	.1659	388	.1637
Other	8054	.4195	902	.4227	995	.4198
Total	19200		2134		2370	

In order to obtain more information about the selected documents and to fine-tune the hyperparameters of the NLP models, an analysis of the text length was performed. Length is measured by the number of tokens present in the text after removing common separators and punctuation marks that do not provide context. Figure 1 shows the number of documents per number of tokens when 100%, 95%, 90% and 75% of the corpus documents are selected. It is observed

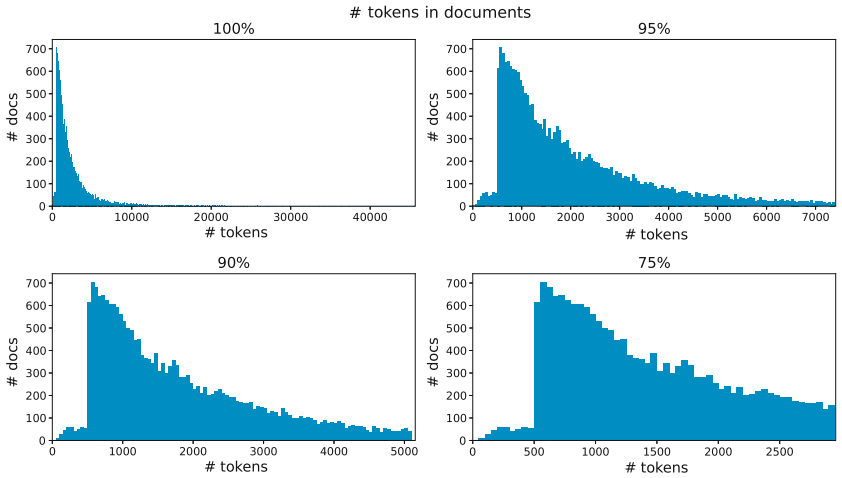

Fig. 1. Distribution of the number of documents by number of tokens when 100%, 95%, 90% and 75% of the corpus documents are considered.

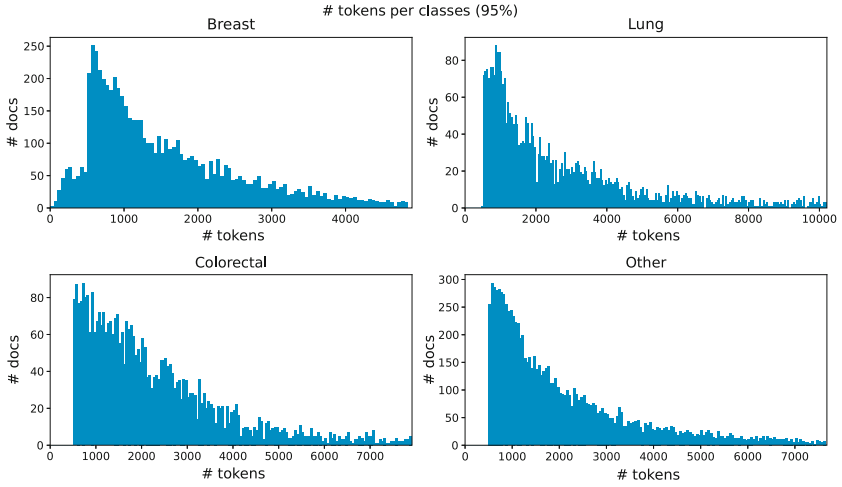

Fig. 2. Distribution of the number of documents by number of tokens for each neoplasm (breast, lung, colorectal, other) when 95% of the corpus documents are considered.

that the maximum number of tokens present in a document is over 40,000, but most documents have less than 10,000 tokens. This is more noticeable when 95% and 90% of the documents are selected. 95% of the documents have less than 7,000 tokens, while 90% have less than 5,000 tokens. Finally, 75% have less than 3,000 tokens, but this clearly implies a lower performance of the NLP models by omitting too large a number of tokens.

When analysing the number of tokens per document belonging to each of the neoplasm classes, 95% of the selected documents is the optimal percentage.

Figure 2 shows these results for breast, lung, colorectal and other neoplasms. There is a clear difference between the number of maximum tokens present in breast and lung neoplasm documents compared to the rest. The 95% of lung documents have less than 10,000 tokens with an average of 3,259.93 tokens, while breast documents are drastically shorter, with less than 5,000 tokens with an average of 1,820.69. Both colorectal neoplasm and other neoplasm documents have a number of tokens less than 7,000 with an average of 2,767.51 and 2,545.71 tokens respectively. Because of these differences, and to ensure that the models are not biased in their choice of neoplasm identified in the EHR solely by the number of tokens present, the hyperparameter fine-tuning with respect to the number of features were set between 5,000 and 7,000 tokens.

3 Methods

This section presents the distinct NLP methodologies developed in this study to tackle the neoplasm type extraction from real-world EHRs in Spanish. The NPL methodologies addressed include ML models, such as NB, SVM and XGBoost; embedding models used in a supervised task; and DL recurrent models, used in conjunction with CNNs and embedding models, such as Word2Vec or fastText.

3.1 Bag-of-Words and Machine Learning Supervised Methods

ML algorithms have been widely used for text processing. However, these methods cannot deal directly with raw text/symbol sequences of variable length, but with numerical feature vectors of fixed size. For this reason, it is necessary to perform a pre-processing of the data for its treatment. Bag-of-Words (BoW) [31] is the most commonly used method for this purpose. BoW transforms documents into a reduced and simplified representation based on criteria such as word frequency, ignoring the order of words and context. BoW creates a dictionary as large as all the different words present in the corpus or limited to the most important or frequent. This dictionary, also known as the vocabulary, is used to vectorise the document, so that the vocabulary is represented as a vector in which each feature is a word stored in it, and its value depends on whether this word occurs in the text and on the criterion chosen.

Count vector and tf-idf are the most common criteria. Count vector is the simplest criterion, where each value associated with a token/word is the number of occurrences of that token, also called term frequency (tf), in the text unit. Term frequency - inverse document frequency (tf-idf) is the combination of tf and inverse document frequency (idf) [25]. The idf assigns a higher weight to words with high or low frequency terms in the document. Thus, the tf-idf value increases in proportion to the number of times a word/token occurs in the document, but is offset by the frequency of the word in the document collection, which reduces the effect of implicitly common words in the corpus.

Once the documents have been vectorised with BoW using the tf-idf criterion, we applied the ML models. The most common ML models considered for this

task are NB, SVM, DTs and RF. On one hand, NB classifiers are known for being simple but efficient algorithms. NB classifiers make the naive assumption that all features belonging to the same class are independent and contribute equally to the categorisation result. This assumption is generally not true in real-world situations. NB then calculates the conditional probability of each class, given a set of features, using Bayes' theorem. On the other hand, SVMs aim to obtain a hyperplane that performs a partitioning of the data. For this purpose, SVM maps the input points onto a higher dimensional feature space, so that the decision boundary maximises the margin between the different classes, thereby clustering them. Prediction involves classifying a sample according to the closest cluster.

Finally, RF and DTs were used with eXtreme Gradient Boosting (XGBoost) [4]. XGBoost is a supervised learning method based on DTs and improves on other methods such as RF and Gradient Boosting by using multiple optimisation methods. Like RF, XGBoost uses ensembles of DTs, but differs in using an additive strategy. In this way, each DT is trained by taking into account the residuals, the difference between the predicted value and the observed value, obtained from the previous DT and optimised using regularisation, pruning and parallel learning methods. Each subsequent DT learns from the previous trees and is not given the same weight. In the prediction process, the model output class is calculated by adding the output of each tree multiplied by a learning rate to the initial prediction. The Python package scikit-learn [20] was used to implement the ML and BoW methods.

3.2 Word Embedding and Recurrent Neural Models

The development of more complex models in recent years, has led to the introduction of new methods, such as word embedding, which incorporate concepts such as similarity of words and part-of-speech tagging. Word embedding is a learning technique where each word or phrase in the vocabulary is mapped to an N-dimensional vector of real numbers. Word2Vec (W2V) and fastText are two of the most commonly used methods for translating n-grams into understandable input for RNNs models.

W2V model is based on maximum likelihood and conditional probabilities, which can be seen as the probability of a word given some of the surrounding words in the corpus. The distance between two words is very close if they can substitute each other given the context. The general training of a W2V model considers a fixed window to observe a word and the rest around the word within the sentence to obtain a context. Within W2V there are two variations, the continuous Bag-of-Words (CBOW) [16] and the skip-gram [17]. The CBOW model assumes that a word is generated as a function of the words surrounding it in the text sequence. That is, the model considers the conditional probability of generating a core word based on the context words present in the window. Thus, each word in the dictionary has two vectors, one when it is used as a centre word and one when it is used as a context word. The vector associated with the context word is generally used as a representation of the document tokens in the CBOW model. The skip-gram model is similar to the CBOW model, but

assumes that a core word can be used to generate the surrounding words in a text sequence. In contrast to CBOW, in skip-gram the core word is usually used as the representation of a word in the transformation of the text unit. It is important to note that both skip-gram and CBOW are self-supervised models, since the supervision comes from untagged data.

The use of n-grams is the main difference between fastText [3] and W2V. FastText operates at a granular level, where words are represented by the sum of the vectors of n-gram frames, whereas Word2Vec only learns vectors for whole words found in the training corpus. This model can produce better vector representations for rare and out-of-vocabulary words because it takes into account the shared parameters of subwords among words with similar structures. The fastText model can also be applied as a reliable text classification algorithm [9]. For this purpose, the structure of the model consists of a hidden layer and an output layer and is quite similar to that of CBOW. The fastText input consists of a sentence with embedded n-gram features averaged as a feature representation of the text. Since the number of n-grams is greater than the number of words, it is impossible to store them all. FastText divides all n-grams into buckets using the hashing track approach, so that they can share an embedding vector. The input layer is summed with the hidden layer, averaged and multiplied by a weight matrix. To produce the output of the model, the hidden layer is then multiplied by another matrix of weights. In order to apply the fastText and W2V methods as embedding models, the Python package gensim [21] was used.

The input stream processed by the embedding is comprehensible to RNNs, which are DL networks specially designed with interconnected units that form an internal memory to deal with problems of temporal structure. RNNs include the GRU [5] and the LSTM [8] networks. The main difference between the two networks is the number and functionality of their internal units. GRU consists of two gates: a reset gate, which determines the amount of past knowledge transferred to the current state; and an update gate, which determines the amount of new information added to the current state. LSTM network consists of three gates: the input gate, which derives the values used to modify the memory; the forget gate, which derives the features to discard; and the output gate, which determines the output based on the input and block memory.

In order to learn the future and past context of the input sequences, both recurrent networks can be structured to form a bidirectional model [12]. This model consists of two layers of recurrent units, GRU or LSTM. To learn the past context, one layer processes the forward sequence based on the current input and the state of the previous hidden unit. To learn the future context, the other layer processes the backward sequence based on the current input and the state of the subsequent hidden unit. The outputs of both recurrent layers are concatenated to feed the other network layers. The model implemented in this study includes a dense layer with ReLU activation function at the output of the recurrent bidirectional layer and a final dense layer with softmax activation function to infer the neoplasm assigned to the document. In addition, to further sensitise the network to the context of the sequence, two bidirectional recurrent layers (2-BiLSTM or 2-BiGRU) can be coupled.

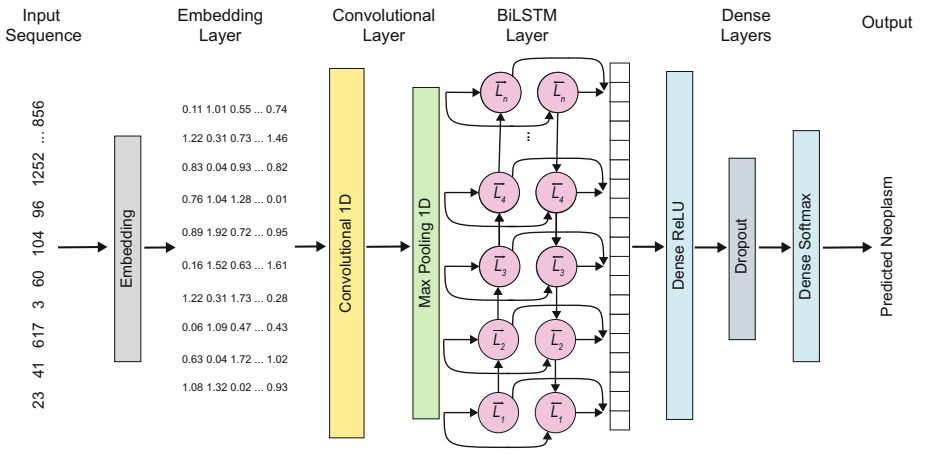

Fig. 3. The CNN + BiLSTM model structure for the text classification system.

In addition, it is possible to add a convolutional layer at the top of these recurrent models [23]. The purpose of this layer is to capture sequence information and reduce input dimensionality in order to feed the recurrent layers. The window of the convolution layer moves across the text representation to extract features, generating sequences that capture the syntax and semantics of the text. The diagram of a BiLSTM model with a convolutional layer and an embedding layer (C-BiRNN) is shown in Fig. 3. The initial sequence is processed by a tokenizer and an embedding model, transforming it into a sequence of tokens with word index values. This input sequence is fed to the embedding, which can be pre-trained or not, to obtain the word vectors that feed the 1D convolutional layer, including a max-pooling. The sequence is fed to the bidirectional recurrent layer and the result is concatenated to feed a dense layer with dropout. Finally, the output layer infers the neoplasm associated with the sequence. Tensorflow [1] package was used to implement these models.

4 Experiments

The experiments performed and the evaluation metrics used in this study are presented in this section. For the experiments, a stratified division of the data into training, validation and test sets is carried out with the data already preprocessed. Table 1 in Sect. 2 describes the result of this division, the number of documents in each set and the corresponding number of neoplasms. In view of the division of data, NLP methods (including BoW, W2V and fastText) are trained using the training set. The validation set is used in a hyperparameter fine-tuning process to achieve maximum classification performance, while the final prediction is performed on the test set. Thus, complete independence is maintained between training, parameter selection and the final prediction performance.

The metrics precision, recall and F1-score were used to evaluate the clinical text classification methods studied. The evaluation metrics are calculated using the true positive (TP), false positive (FP) and false negative (FN) values of the confusion matrix. The precision metric indicates the ratio of correctly predicted documents belonging to a neoplasm to the total number of positively predicted documents. Meanwhile, recall, also known as sensitivity or true positive rate (TPR), is the ratio of correctly predicted documents belonging to a neoplasm to the total number of documents of the actual neoplasm. Equation 1 formally defines the calculation of both metrics. The F_1 score is the harmonic mean of precision and sensitivity (Eq. 2) and provides a reliable measure of the prediction performance achieved in problems where sensitivity is important.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{2}$$

In addition, considering that we are studying a multi-class problem, the micro- and macro-average were considered [24]. Once the evaluation metrics were defined, the micro-average was calculated by summing the individual TP, FP and FN provided by the prediction for the different classes, and then calculating the precision, recall and F1-measure metrics. In contrast, the macro-average simply performs the average of each of the computed metrics. Since the neoplasm types present in Galén's corpus are slightly unbalanced, the macro-average evaluation is given greater weight.

5 Results

The experimentation process proposed above was followed, and Table 2 shows the neoplasm classification results, precision, recall and F1-score values achieved in macro-average (ovr-average) and micro-average evaluation. The best values achieved are shown in bold, while the second best values are shown in italics. Two main conclusion can be drawn from the results described in Table 2. On the one hand, according to the F1-score metric, the methods based on the application of BoW and ML used in this work obtain, on average, a lower performance. The SVM method is the only one that outperforms the others, surpassing the performance obtained with some RNNs. SVM obtains a micro- and macro-average F1-score of 0.9814 and 0.9788, respectively. On the other hand, RNN-based models with convolutional layers and fastText embedding outperform the other methods in extraction of neoplasm type. It is important to note that the same architectures outperform their versions with W2V embedding, and that RNNs with GRU units outperform LSTM units. Among all the methods, the best performance is achieved by the 2-BiGRU model when a pre-trained CBOW fastText embedding model is applied, obtaining a micro- and macro-average F1-score of 0.9840 and 0.9821, respectively.

Table 2. Micro- and Macro-averaged metrics computed on test set. For each evaluation strategy, precision (P), recall (R) and F1-score (F1) metrics are computed.

Model	micro			macro		
	P	R	F1	P	R	F1
Naive Bayes	.9683	.9654	.9668	.9618	.9652	.9634
SVM	.9814	.9814	.9814	.9768	.9808	.9788
XGBoost (DT)	.9561	.9561	.9561	.9530	.9512	.9520
XGBoost (RF)	.9776	.9776	.9776	.9751	.9749	.9750
fastText	.9814	.9793	.9804	.9790	.9772	.9781
W2V + C-BiLSTM	.9780	.9776	.9778	.9746	.9761	.9753
W2V + C-2-BiLSTM	.9768	.9759	.9764	.9737	.9717	.9727
W2V + C-BiGRU	.9823	.9819	.9821	.9805	.9778	.9791
W2V + C-2-BiGRU	.9789	.9789	.9789	.9767	.9760	.9763
FT + C-BiLSTM	.9827	.9823	.9825	.9791	*.9812*	.9801
FT + C-2-BiLSTM	.9840	*.9831*	*.9835*	**.9837**	.9788	.9812
FT + C-BiGRU	**.9848**	*.9831*	**.9840**	**.9837**	.9797	*.9817*
FT + C-2-BiGRU	*.9844*	**.9835**	**.9840**	*.9806*	**.9838**	**.9821**

Overall, the results in Table 2 do not show a clear effectiveness of obtaining the context of the sequences compared to the other methodologies in text classification. There are two possible reasons for the observed high performance of ML methods with BoW, which do not capture context. One could be the inference of the neoplasm presented in the document by key clinical concepts that are different for each neoplasm, such as the mention of a specific diagnostic test, for example a mammogram in the case of breast neoplasms. This is to be expected and is perfectly acceptable. However, another reason for this behaviour could be the inference of neoplasms by the presence of non-clinical concepts, such as the attending oncology specialist or the medical centre mentioned in the history. The inference based on the presence of these concepts is not desired, as the use of these pre-trained models in medical centres in other regions could lead to errors.

Table 3. Metrics for each neoplasm obtained by the convolutional 2-Bidirectional GRU with fastText embedding (FT + C-2-BiGRU) system on the test set.

Neoplasm	P	R	F1	Support
Breast	.9954	.9969	.9962	649
Lung	.9622	.9793	.9707	338
Colorectal	.9769	.9820	.9794	388
Other	.9878	.9769	.9823	995
micro-avg	.9844	.9835	.9840	2370
macro-avg	.9806	.9838	.9821	2370

With the aim of conducting a thorough analysis of the performance of the convolutional 2-Bidirectional GRU model with pre-trained fastText embedding (FT + C-2-BiGRU), which achieves the best neoplasm extraction results, Table 3 shows the metrics obtained for each of the neoplasms considered separately. The number of texts associated with each neoplasm in the test set is also shown, for a better evaluation of the obtained metrics. The C-2-BiGRU model performs particularly well in the classification of breast neoplasms, with an F1-score of 0.9962 and a recall of 0.9969. As the support value shows, this is the most common independent neoplasm in the Galén corpus (649 documents), which explains the higher performance. For lung and colorectal neoplasms, the results obtained are acceptable, especially for recall, where they outperform precision with values of 0.9793 and 0.9820 respectively. Apart from the F1-score, recall is the most important metric in diagnostic support systems. Taking into account that the category of other neoplasms includes neoplasms that may be related to the previous ones, especially lung and colorectal (e.g. due to the diagnostic tools and the regions of the body in which they are performed), this category performs slightly better, with a value of 0.9823 F1-score.

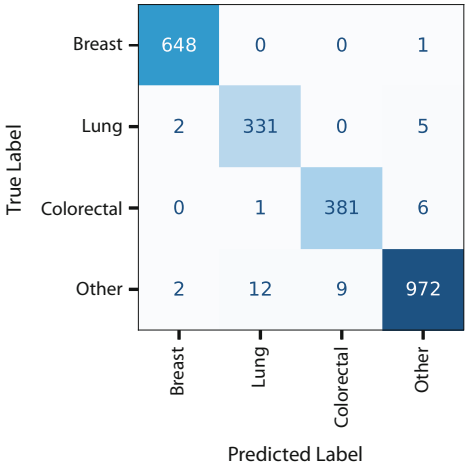

Fig. 4. Confusion matrix obtained with FT + C-2-BiGRU method on the test set.

Finally, Fig. 4 shows the confusion matrix obtained by the C-2-BiGRU model with fastText on the test set. As we can see from the matrix, the number of misclassifications for this particular complex text classification task is low. The total number of misclassified documents is 38 out of 2370 documents, giving an accuracy of 0.984.

6 Conclusions

In this paper we have addressed the problem of the extraction of neoplasm type from real-world clinical documents in Spanish. For this purpose, we have elaborated a corpus of 23,704 medical cases annotated with the neoplasm presented by the patient, obtained from Galén [22]. The performance of ML and BoW-methods and RNN-based models applied to the text classification task has been analysed. The results obtained show that, on the one hand, BoW-based methods achieve similar results to those that consider the context of the sequences. This is probably caused by two factors: the presence of clinical concepts related to neoplasms, such as mammography and breast cancer, or the presence of non-clinical concepts, such as the names of clinical specialists who treat certain neoplasms. As the corpus was obtained from a small group of oncology centres, further analysis is needed to refute this second idea. On the other hand, bidirectional RNNs with a convolutional layer and pre-trained fastText embedding outperform the others methodologies for neoplasm type extraction. Among these RNN-based systems, the best performance is obtained by the C-2-BiGRU model with fastText, with macro-averaged precision, recall and F1-score of 0.9806, 0.9838 and 0.9821, respectively. In terms of neoplasm types, the neoplasms best classified by the C-2-BiGRU model with fastText are breast neoplasms, followed by other neoplasms (includes head/neck, non-Hodgkin's lymphoma, thyroid, stomach/esophagus and other), colorectal and lung neoplasms.

In future work, we will investigate the extraction of the type of neoplasm from the selected Galén corpus, but without obfuscation. In order to preserve the privacy of the data, we will perform a de-identification process, where private concepts (names, identifiers, medical centres, locations, etc.) will be randomly replaced by others, so that the context is preserved, but avoiding that NLP models learn from non-clinical concepts. Finally, an attempt will be made to validate the developed methodology on external real-world corpora from other Spanish medical centres, given the promising results of this work.

Acknowledgements. The authors acknowledge the support from the Ministerio de Ciencia e Innovación (MICINN) under project PID2020-116898RB-I00, from Universidad de Málaga and Junta de Andalucía through grants UMA20-FEDERJA-045 and PYC20-046-UMA (all including FEDER funds), and from the Malaga-Pfizer consortium for AI research in Cancer - MAPIC.

References

1. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous systems (2015). https://www.tensorflow.org/
2. Baker, S., Korhonen, A., Pyysalo, S.: Cancer hallmark text classification using convolutional neural networks. In: Proceedings of the 5th Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), pp. 1–9 (2016)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017). https://doi.org/10.1162/tacl_a_00051

4. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785
5. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Workshop on Deep Learning, December 2014 (2014)
6. Garla, V., Taylor, C., Brandt, C.: Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. J. Biomed. Inform. **46**(5), 869–875 (2013). https://doi.org/10.1016/j.jbi.2013.06.014
7. Hadi, W., Al-Radaideh, Q.A., Alhawari, S.: Integrating associative rule-based classification with naïve bayes for text classification. Appl. Soft Comput. **69**, 344–356 (2018). https://doi.org/10.1016/j.asoc.2018.04.056
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
9. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
10. Kasthurirathne, S.N., et al.: Toward better public health reporting using existing off the shelf approaches: the value of medical dictionaries in automated cancer detection using plaintext medical data. J. Biomed. Inform. **69**, 160–176 (2017). https://doi.org/10.1016/j.jbi.2016.01.008
11. Khadhraoui, M., Bellaaj, H., Ammar, M.B., Hamam, H., Jmaiel, M.: Survey of BERT-base models for scientific text classification: COVID-19 case study. Appl. Sci. **12**(6), 2891 (2022). https://doi.org/10.3390/app12062891
12. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)
13. Liu, G., Guo, J.: Bidirectional LSTM with attention mechanism and convolutional layer for text classification. Neurocomputing **337**, 325–338 (2019). https://doi.org/10.1016/j.neucom.2019.01.078
14. López-García, G., Jerez, J.M., Ribelles, N., Alba, E., Veredas, F.J.: Detection of tumor morphology mentions in clinical reports in Spanish using transformers. In: Rojas, I., Joya, G., Català, A. (eds.) IWANN 2021. LNCS, vol. 12861, pp. 24–35. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85030-2_3
15. Mendonça, E.A., Haas, J., Shagina, L., Larson, E., Friedman, C.: Extracting information on pneumonia in infants using natural language processing of radiology reports. J. Biomed. Inform. **38**(4), 314–321 (2005). https://doi.org/10.1016/j.jbi.2005.02.003
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, vol. 26 (2013)
18. Moschitti, A., Basili, R.: Complex linguistic features for text classification: a comprehensive study. In: McDonald, S., Tait, J. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 181–196. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24752-4_14
19. Nguyen, A.N., et al.: Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J. Am. Med. Inform. Assoc. **17**(4), 440–445 (2010). https://doi.org/10.1136/jamia.2010.003707
20. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

21. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, pp. 45–50. ELRA (2010). http://is.muni.cz/publication/884893/en
22. Ribelles, N., et al.: Galén: Sistema de información para la gestión y coordinación de procesos en un servicio de oncología. RevistaeSalud **6**(21), 1–12 (2010)
23. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
24. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. **45**(4), 427–437 (2009). https://doi.org/10.1016/j.ipm.2009.03.002
25. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. **28**(1), 11–21 (1972). https://doi.org/10.1108/eb026526
26. St-Maurice, J., Kuo, M.H., Gooch, P.: A proof of concept for assessing emergency room use with primary care data and natural language processing. Methods Inf. Med. **52**(01), 33–42 (2013). https://doi.org/10.3414/ME12-01-0012
27. Urda, D., Ribelles, N., Subirats, J.L., Franco, L., Alba, E., Jerez, J.M.: Addressing critical issues in the development of an oncology information system. Int. J. Med. Inform. **82**(5), 398–407 (2013). https://doi.org/10.1016/j.ijmedinf.2012.08.001
28. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
29. Venkataraman, G.R., et al.: Fastag: automatic text classification of unstructured medical narratives. PLoS ONE **15**(6), e0234647 (2020). https://doi.org/10.1371/journal.pone.0234647
30. Vítores, D.F.: El español: una lengua viva. Informe 2019. Instituto Cervantes (2019). https://www.cervantes.es/imagenes/File/espanol_lengua_viva_2019.pdf
31. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 977–984 (2006). https://doi.org/10.1145/1143844.1143967
32. Wang, R., Li, Z., Cao, J., Chen, T., Wang, L.: Convolutional recurrent neural networks for text classification. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2019). https://doi.org/10.1109/ijcnn.2019.8852406
33. Yao, L., Mao, C., Luo, Y.: Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med. Inform. Decis. Mak. **19**(3), 31–39 (2019). https://doi.org/10.1186/s12911-019-0781-4