Aiding Clinical Triage with Text Classification

Rute Veladas[1](✉), Hua Yang[1,4], Paulo Quaresma[1](✉), Teresa Gonçalves[1](✉), Renata Vieira[2], Cátia Sousa Pinto[3], João Pedro Martins[3], João Oliveira[3], and Maria Cortes Ferreira[3]

[1] Department of Informatics, University of Évora, Évora, Portugal
{rgv,huayang,pq,tcg}@uevora.pt
[2] CIDEHUS, University of Évora, Évora, Portugal
renatav@uevora.pt
[3] Serviços Partilhados do Ministério da Saúde, Lisboa, Portugal
{catia.pinto,joao.martins,joao.oliveira,
ricardo.vicente,maria.cortes}@spms.min-saude.pt
[4] Zhongyuan University of Technology, Zhenghou, China

Abstract. SNS24 is a telephone service for triage, counselling, and referral service provided by the Portuguese National Health Service. Currently, following the predefined 59 Clinical Pathways, the selection of the most appropriate one is manually done by nurses. This paper presents a study on using automatic text classification to aid on the clinical pathway selection. The experiments were carried out on 3 months calls data containing 269,669 records and a selection of the best combination of ten text representations and four machine learning algorithm was pursued by building 40 different models. Then, fine-tuning of the algorithm parameters and the text embedding model were performed achieving a final accuracy of 78.80% and F1 of 78.45%. The best setup was then used to calculate the accuracy of the top-3 and top-5 most probable clinical pathways, reaching values of 94.10% and 96.82%, respectively. These results suggest that using a machine learning approach to aid the clinical triage in phone call services is effective and promising.

Keywords: Machine learning · Text classification · Clinical triage · SNS24

1 Introduction

According to the EuroHealth Consumer Index, 17 European countries had some form of telephone clinical triage by 2018 [3]. SNS24 is a telephone service for triage, counselling and referral service provided by the Portuguese National Health Service. In 2018 SNS24 answered more than 1 million calls with an average duration of 7–8 min. Being of national scope, this is a service that promotes equity in the access to health care.

G. Marreiros et al. (Eds.): EPIA 2021, LNAI 12981, pp. 83–96, 2021.
https://doi.org/10.1007/978-3-030-86230-5_7

The SNS24 telephone service is provided by nurses and follows predefined clinical pathways. Triage is based on the selection of a clinical pathway by a nurse, considering the citizen's self reported symptoms and signs as well as relevant information provided on medical history. The choice of the most appropriate pathway is extremely important and relevant since it should ensure **high safety** (not failing to identify situations that require urgent medical contact) and have high **discriminatory capability** (do not send low clinical risk situations to Hospitals' Emergency). This is a quite complex problem because there are 59 possible clinical pathways, with five possible final referrals: self-care, observation at primary health care center, observation in hospital emergency, transference to the National Medical Emergency Institute or to the Anti-Poison Information Center.

During each shift, nurses log on to the platform and start answering incoming calls from a single queue. Each call begins with screening questions for emergency situations, in which case the call is redirected to National Medical Emergency Institute. For non-emergencies, the call ensures identification confirmation, specific protocol choice (that are elicited by a keyword search box) and free-text records for documentation purposes. Figure 1 illustrates this process.

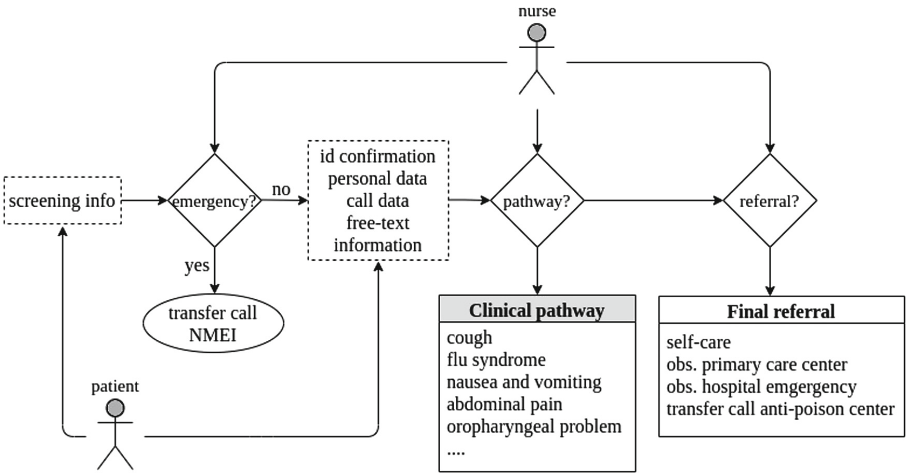

Fig. 1. SNS24 triage process

For each possible referral, there are different situations that need to be assessed. For instance, in a hospital ER there are several possible Manchester triage decisions and they should be analyzed and co-related with the SNS24 referral decision. Moreover, age and sex of the user and even other variables, such as date and time (e.g. day versus night situations) and clinical experience of the triage nurse should also be considered in this process.

Although promoting equity in access to health, SNS24 service can still accommodate many improvements such as allow a better and faster interaction between the citizens and SNS24 and decrease the duration of calls aiming to increase the quantity of handled and further improve the availability of the SNS24 service. Obviously, an improvement in the quality of the algorithm selection and a decrease of the average duration of the phone calls will have a major impact in the SNS24 service to citizens.

Therefore, it is important to support nurses in the selection of the most appropriate clinical pathway and optimize them through the analysis of post referral diagnosis made at primary care units and hospitals. This will improve the discriminatory capacity and clinical safety of the correspondent referrals, and allow SNS24 to improve quality of the existing services.

In this work, we focus on the task of selecting the most appropriate clinical pathway for clinical triage. The paper is organized as follows, in Sect. 2 we review related work; in Sect. 3 we describe the materials and methods used; in Sect. 5 we describe the experiments performed and analyse the results; and finally in Sect. 6 we draw conclusions.

2 Related Work

It is known that the use of clinical decision support systems improves the quality of telephone triage service [21] and the performance of care [9]. The current rate of adherence by citizens to recommendations is partial and reported as moderate, where rates of adherence to self-care, primary health care and hospital emergencies are 77.5%, 64.6% and 68.6% in Australia, respectively [30]. However, although these conclusions have not been consolidated [12], when the recommendations of these guidelines are followed, more suitable referrals to the Emergency Service are obtained [7] and a cost reduction can be achieved [20].

Machine Learning and Natural Language Processing techniques have been increasingly applied in clinical decision support systems, and there is a growing effort in applying these techniques to clinical narrations [11,14,17,19,25,26]. There is a wide variety of paradigms for classification problems (e.g. linear, probabilistic, neural networks) and, for each, there are several algorithms [1]. However, there is no classification algorithm that can be considered the best for all problems [8].

In the last years many conferences and evaluation tasks have focused on clinical decision support problems [10] and the overall performance of the best systems has improved with the use of the new machine learning approaches, such as Support Vector Machines and Deep Learning architectures [17,26,28]. Recent deep learning approaches typically use architectures based on bidirectional LSTM with attention mechanisms having as input word embedding vectors [25]. Current ongoing research tries to create hybrid systems, integrating pure Machine Learning algorithms with linguistic information, such as part-of-speech (POS) and syntactic and semantic information [32].

Mascio *et al.* [15] compared various word representations and classification algorithms for clinical text classification tasks. They experimented on four datasets and found that traditional word embeddings (Word2Vec, GloVe and

FastText) could achieve or exceed the performance of the ones based on contextual embeddings (BERT) when using the neural-network based approaches, while using traditional machine learning approaches (SVM), the contextual words embeddings achieved better performance. Topaz *et al.* [29] studied patients with high risk of hospitalization or emergency visits using clinical notes taken during home health care episodes. This study was experimented on a database which included 727,676 documents for 112,237 episodes and 89,459 unique patients; they used text mining and machine learning algorithms for the prediction, and feature selection techniques to find risk factors, and found that the using a Random Forest algorithm achieved the best performance with an F-measure of 0.83. Flores *et al.* [6] studied how to achieve the specified performance in biomedical text classification while reducing the number of labeled documents. They compared an active learning approach with Support Vector Machines, Naïve Bayes and a classifier based on Bidirectional Encoder Representations from Transformers (BERT) and experimented on three datasets with biomedical information in Spanish; the active learning approach obtained a AUC (area under the learning curve) performance greater than 85% in all cases. Mullenbach *et al.* [18] used a convolutional network and an attention mechanism to predict medical codes from clinical texts. Their method achieved a precision of 71% and a micro-F1 of 54%.

These related works give an idea of the technologies employed in the area. However there are no works to perform actual comparisons since the experimental setup, problem definition and datasets involved are particular for the demand of SNS24. Instead, we compared simple and more complex text representations as well as some of the classical text classification techniques adopted in these previous works for our problem.

3 Materials and Methods

This section details the materials and methods selected for the execution of the task, containing the description of the data set, the selection of the attributes and the set of experiments selected.

3.1 Available Data

The study protocol was approved by the competent ethics committee and the anonymized data was provided by SPMS (Serviços Partilhados do Ministério da Saúde). It has a total of 269,663 records with 18 attributes, corresponding to information collected during 3 months of calls received by the SNS24 phone-line (from January to March 2018). It includes personal data (age, gender, encrypted primary care unit) and call data (start and end date/time, initial intention, comments, contact reason, clinical pathway and final disposition, between others). The contact reason and comments are free text written in Portuguese by the technician who answered the respective call; the remaining are nominal attributes (except dates).

3.2 Task

With the available data, the task of selecting the most appropriate clinical pathway can be framed as a supervised multi-class classification problem where the attribute "Clinical pathway" is the class aiming to be predicted.

3.3 Dataset

From the 59 possible clinical pathways, in the provided data there was only examples of 53. The proportion of observations per class is diverse ranging from 14.006% for "Tosse/Cough" to 0.001% for "Problemas por calor/Heat-related problems"; 5 clinical pathways have proportions above 5% and 27 have less than 1%. Table 1 presents the five clinical pathways with more and less observations.

Table 1. The five clinical pathways with more and less observations.

Clinical pathway (class)	Examples	%
Tosse/Cough	37930	14.066
Síndrome Gripal/Flu syndrome	34266	12.707
Prob. por náuseas e vómitos/Nausea and vomiting...	14453	5.360
Dor abdominal/Abdominal pain	14382	5.333
Problema da orofaringe/Oropharyngeal problem	13503	5.007
Problemas no cotovelo/Elbow problems	137	0.051
Problemas por sarampo/Measles related problems	98	0.036
Prob. adaptaçño situaçño de crise/... crisis situation	60	0.022
NA	5	0.002
Problemas por calor/Heat problems	4	0.001

The available attributes were analysed and the first one selected for the experiments was "Contact Reason", a medium length free text attribute containing simple and straight-forward information about the patient's problem. It has a total of 31,417 distinct words with each value composed by an average number of 8.15 words (and standard deviation of 3.64). Table 2 presents a few examples for the "Contact Reason".

For building the dataset, clinical pathways with less than 50 instances were removed from the original data, resulting in a dataset with 269,654 instances.

3.4 Text Representation

The "Contact Reason" text was pre-processed with simple word count and Term Frequency–Inverse Document Frequency (TF-IDF) methods, which determine the importance of a word in a corpus, and also with word embedding models, which map the words into a low-dimensional continuous space encoding their semantic and syntactic information (by assuming that words in similar context should have similar meaning) [13].

Table 2. Examples of the contact reason field

Febre desde esta manhã Fever since this morning
Dor no ouvido esquerdo com tonturas associadas por 4 dias Left ear pain with associated dizziness for 4 days
Tosse produtiva, congestño nasal e febre há 7 dias Productive cough, nasal congestion and fever for the last 7 days
Dor de garganta desde Sexta e 38.7 °C de febre Sore throat since Friday and 38.7 °C fever
Dor de cabeça, mialgias e tosse com expetoraçño verde Headache, myalgias and cough with green expectoration
Laceraçño do couro cabeludo há 5 minutos Scalp laceration 5 min ago
Dor no pescoço e garganta após biópsia à tiróide há 18h Pain in neck and throat after thyroid biopsy 18 h ago

3.5 Experiments

A first set of experiments was done on the validation set to determine the combination of the machine learning algorithm and the text representation that produced the best results over the dataset. The algorithms tested were Support Vector Machines (with linear and rbf kernels), Random Forest and Multinomial Naïve Bayes; the "Contact Reason" text was processed to build the following text representations: word n-grams ($n \in \{1, 2, 3, 4\}$) using word counts and TF-IDF and embeddings using BERT [5] and Flair [2] models. Both Flair and BERT used pre-trained models publicly available for the Portuguese language.

After selecting the best combination of representation and algorithm, by using statistical McNemar tests to compare the results, a fine-tuning of the embedding model was performed and tested on both validation and test sets. Finally the performance of the generated model using the most probable class along with the three (top-3) and five (top-5) most probable ones were calculated. This was done since the purpose of the classification is to help nurses to find the clinical pathway and offering the most probable ones may help in their decision.

3.6 Experimental Setup

For developing the models Python (v3.7.9) along with scikit-learn (v0.23.2), Transformers[1] (v3.4.0) and Flair[2] (v0.6.1) were used.

[1] https://huggingface.co/transformers/v3.4.0/.

[2] https://github.com/flairNLP/flair.

The language models employed were FlairBBP[3] and BERT Large (BERTimbau[4]). Flair embeddings are based on the concept of contextual string embeddings which are used for sequence labelling. The FlairBBP language model was developed on the basis of a raw text corpus of 4.9 billion words from contemporary Portuguese texts. It was previously evaluated for the NLP task of named entity recognition [22] and also in specific domains like geoscience [4], law [24], and health [23]. BERTimbau was trained on the BrWaC (Brazilian Web as Corpus), a large Portuguese corpus [31] and evaluated on several NLP tasks [27].

A stratified split of the dataset into train, validation and test sets was made with a distribution of 64%, 16% and 20%, respectively. The first set of experiments (Sect. 4.1) was evaluated over the full split of validation set with the test split being used for the final evaluation of the best model (Sect. 4.2).

The models were evaluated using accuracy and weighted average of F1-measure (we also present the weighted precision and recall values of each experiment). To support the choice of the best model(s) McNemar tests [16] were performed with a level of significance $\alpha = 0.05$; in the results' tables, the significantly best performing models are presented in bold-face.

4 Results

This section presents the results obtained for each set of experiments done: (1) selection of the "best" combination of the algorithm and representation, (2) fine-tune of the embedding model and (3) calculation of the selected model performance using the most probable class along with the three (top-3) and five (top-5) most probable ones.

4.1 Find the "Best" Algorithm and Representation

This stage corresponds to the development of the models previously mentioned using "Contact Reason" attribute (see Sub-sect. 3.5), totaling 40 models using 4 different machine learning algorithms and 10 different text representations. The performance results were calculated over the validation set and are organized by the machine learning algorithm.

After obtaining the 40 models with default parameters for each algorithm, a fine-tuning of parameters was performed for the best combination of algorithm and representation.

Support Vector Machines. Table 3 and Table 4 present the results using the Support Vector Machine algorithm using linear and RBF kernels, respectively.

When comparing SVMs for the same representation, the linear SVM model always has a better performance with the exception of the uni-grams with TF-IDF and RBF SVM model. It is possible to observe that when using word n-grams the performance decreases when increasing n and that TF-IDF consistently produced better results when compared to using a simple n-gram count.

[3] https://github.com/jneto04/ner-pt#flair-embeddings---flairbbp.

[4] https://github.com/neuralmind-ai/portuguese-bert.

Table 3. Linear SVM: performance for different representations.

		Acc.	Prec.	Rec.	F1
Uni-grams	Count	76.28	76.05	76.28	75.99
	TF-IDF	76.47	76.14	76.47	76.10
Bi-grams	Count	72.03	71.74	72.03	71.72
	TF-IDF	73.34	72.90	73.34	72.88
Tri-grams	Count	61.23	62.59	61.23	60.99
	TF-IDF	62.56	63.46	65.56	62.07
Quadri-grams	Count	43.77	54.17	43.77	44.64
	TF-IDF	44.75	54.78	44.75	45.31
Embeddings	BERT	76.39	76.16	76.39	76.04
	Flair	**77.96**	**77.49**	**77.96**	**77.51**

Table 4. RBF SVM: performance for different representations.

		Acc.	Prec.	Rec.	F1
Uni-grams	Count	76.15	76.89	76.15	76.10
	TF-IDF	**76.97**	**77.28**	**76.97**	**76.83**
Bi-grams	Count	68.07	68.74	68.07	67.40
	TF-IDF	70.27	70.48	70.27	69.60
Tri-grams	Count	51.63	62.12	51.63	51.63
	TF-IDF	55.80	63.72	55.80	55.15
Quadri-grams	Count	32.88	62.59	32.88	32.57
	TF-IDF	34.75	62.11	34.75	36.01
Embeddings	BERT	75.83	75.81	75.84	75.52
	Flair	76.58	76.78	76.59	76.30

Random Forest. The results obtained with Random Forest algorithm are presented in Table 5. As observed, the best performance was also obtained using word uni-grams with with TF-IDF, but SVMs consistently generated better models. The observations made about the n-grams performance and TF-IDF for SVMs are also true for Random Forest.

Table 5. Random Forest: performance for different representations.

		Acc.	Prec.	Rec.	F1
Uni-grams	Count	73.84	73.33	73.84	73.26
	TF-IDF	**74.93**	**74.55**	**74.93**	**74.40**
Bi-grams	Count	66.51	66.14	66.51	65.54
	TF-IDF	68.10	67.70	68.11	67.41
Tri-grams	Count	55.83	58.61	55.83	55.73
	TF-IDF	56.73	58.98	56.73	56.39
Quadri-grams	Count	39.67	58.15	39.67	44.67
	TF-IDF	39.82	57.61	39.82	44.65
Embeddings	BERT	69.39	68.94	69.39	68.16
	Flair	68.36	68.43	68.37	66.96

Multinomial Naïve Bayes. Table 6 presents the results obtained with Multinomial Naïve Bayes algorithm. As can be seen, it under-performs the previous algorithms and, unlike the results previously presented, count produces better results when compared to TF-IDF.

Table 6. Multinomial NB: performance for different representations

		Acc.	Prec.	Rec.	F1
Uni-grams	Count	**66.26**	**67.68**	**66.26**	**63.85**
	TF-IDF	57.83	63.10	57.83	53.59
Bi-grams	Count	60.09	64.85	60.09	57.32
	TF-IDF	52.58	65.96	52.58	49.35
Tri-grams	Count	51.35	62.37	51.35	49.35
	TF-IDF	42.70	66.43	42.70	40.55
Quadri-grams	Count	36.18	60.52	36.18	34.74
	TF-IDF	29.45	62.11	29.45	26.20
Embeddings	BERT	58.73	60.16	58.74	57.73
	Flair	47.78	55.54	47.78	45.77

Parameter Optimization. As can be seen from Tables 3, 4, 5 and 6 the top 3 most performing approaches were: Flair with linear SVM, BERT with linear SVM and uni-grams with TF-IDF and RBF SVM. These models were parameter fine-tuned to maximize the F1-measure. Models were built with values of $C \in \{0.01, 0.1, 1, 10, 100, 1000\}$. For Flair and uni-grams the best model was obtained with $C = 1$ (default value) and $C = 0.1$ for BERT representation. For the RBF kernel with uni-gram and TF-IDF, the γ (gamma) parameter was also fine-tuned with different values but the default parameter (inverse of number of features times attribute variance, $1/(nfeatures * var)$) generated the best model.

Table 7 summarizes the results. The only improvement observed was for BERT using linear SVM, but still being lower than the one obtained with Flair. For this reason Flair with linear SVM combination was selected for pursuing the following experiments.

Table 7. SVM performance summary.

	Acc.	Prec.	Rec.	F1
BERT w/ linear SVM	76.39	76.16	76.39	76.04
BERT w/ linear SVM ($C = 0.1$)	**77.10**	**76.54**	**77.10**	**76.61**
Flair w/ linear SVM	**77.96**	**77.49**	**77.96**	**77.51**
Uni-gram TF-IDF w/ RBF SVM	**76.97**	**77.28**	**76.97**	**76.83**

4.2 Fine-Tuning the Embedding Model

The approach selected for the experiments on this stage was Flair using Linear SVM. Aiming to improve the previously obtained results, the pre-trained Flair embedding model was fine-tuned to be adapted to the clinical domain using a corpus built from the "Contact Reason" texts of the SNS24 dataset.

This new embedding model was used to evaluate the performance over the validation set and also the test. The results can be seen in Table 8. According to the significance tests performed, the Flair fine-tuning produced a significant improvement on the performance over the validation set. When applying this model to the test set, an accuracy of 78.80% and F1 of 78.45% were achieved.

Table 8. SVM performance with fine-tuned Flair model

	Acc.	Prec.	Rec.	F1
Original Flair (validation set)	77.96	77.49	77.96	77.51
Fine-tuned Flair (validation set)	**78.59**	**78.18**	**78.59**	**78.21**
Fine-tuned Flair (test set)	78.80	78.42	78.80	78.45

4.3 Considering the Most Probable Clinical Pathways

This section presents the performance when considering the three (top-3) and five (top-5) most probable classes given by the prediction model. This experiment was done using the fine-tuned Flair model with linear SVM and uni-grams TF-IDF with RBF SVM (the two "best" models from stage 1; see Sub-sect. 4.1.

Accuracy results for the test set are shown in Table 9. Looking at the results, and despite the significant difference between models when using the most probable class (top-1), this is no longer true when considering the top-3 and top-5 classes.

Table 9. Accuracy for the top-1, top-3 and top-5 most probable classes.

	Top-1	Top-3	Top-5
Unigram TF-IDF w/ RBF SVM	76.97	94.08	96.77
Fine-tuned Flair w/ linear SVM	**78.80**	94.10	96.82

5 Discussion

The experiments performed on Sub-sect. 4.1 provided information on the setup of the best model for the problem at hand. The best result obtained was with Flair embeddings and linear SVM, with an accuracy of 77.96% and an F1 of 77.51%. In these experiments it is possible to observe that, for all algorithms using n-grams, as the n increases the performance decreases, so uni-grams generated the best models when using word n-grams; for this setup, the best performance was obtained with TF-IDF and RBF SVM with an accuracy of 76.97% and F1 of 76.83%.

For the Random Forest algorithm, uni-grams with TF-IDF generated the best models but it under-performs SVM for all text representations (accuracy of 74.93% and F1 of 74.40%). For the Multinomial Naïve Bayes, the representation with better results was also uni-grams but with count with an accuracy of 66.26% and F1 of 63.85%; this algorithm under-performs by large SVMs and Random Forests. To finalize the initial experiments a fine-tuning of the SVM parameters was performed for several text representations. A small improvement was obtained using BERT but it was still lower than Flair with linear SVM.

Still pursuing the goal of obtaining the best classification model, the pre-trained Flair model was fine-tuned using a corpus composed by the "Contact Reason" text of the SNS24 dataset to be better adapted to the clinical domain. This fine-tuning provided an improvement of the performance, reaching an accuracy of 78.80% and F1 of 78.45% for the test set.

Finally, the accuracy of two prediction models (fine-tuned Flair with linear SVM and unigram TF-IDF with RBF SVM) was measured calculating the top-3 and top-5 most probable classes. The results showed that, despite the significant difference between both models when using the most probable class, it was no longer true when presenting the top-3 or top-5 most probable classes. Consequently, one can say that uni-grams TF-IDF with RBF SVM would be the final model choice to incorporate in a clinical tool since its computational cost is lower when compared with the Flair representation.

6 Conclusions

SNS24 is a telephone triage service provided by the Portuguese National Health Service, where nurses select the most appropriate clinical pathway given the information self-reported by citizens. This paper proposes to use an automatic text classification approach to aid the SNS 24 clinical triage service.

A group of experiments were conducted on 3 months data containing a total of 269,669 call records. Several machine learning algorithms (SVM with linear and RBF kernel, Random Forest and Multinomial Naïve Bayes) and text representations (TF-IDF and count n-grams and BERT and Flair embeddings) were combined to produce classification models. The experimental results show that a fine-tuned Flair embedding with a linear SVM classification model achieves an accuracy of 78.80% and F1 of 78.45%; additionally accuracies of 94.10% and 96.82% were obtained when using the top-3 and top-5 most probable classes. These results suggest that using Machine Learning is an effective and promising approach to aid the clinical triage of phone call services.

Acknowledgement. This research work was funded by FCT – Fundaçño para Ciência e Tecnologia, I.P, within the project SNS24.Scout.IA: Aplicaçño de Metodologias de Inteligência Artificial e Processamento de Linguagem Natural no Serviço de Triagem, Aconselhamento e Encaminhamento do SNS24 (ref. DSAIPA/AI/0040/2019).

References

1. Aggarwal, C.C., Clustering, C.R.D.: Algorithms and applications (2014)
2. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1638–1649. Association for Computational Linguistics, Santa Fe (2018). https://www.aclweb.org/anthology/C18-1139
3. Björnberg, A., Phang, A.Y.: Euro health consumer index 2018 report. In: Health Consumer Powerhouse Euro Health Consumer Index, pp. 1–90 (2019)
4. Consoli, B.S., Santos, J., Gomes, D., Cordeiro, F., Vieira, R., Moreira, V.: Embeddings for named entity recognition in geoscience Portuguese literature. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 4625–4630 (2020)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding (2018)
6. Flores, C.A., Figueroa, R.L., Pezoa, J.E.: Active learning for biomedical text classification based on automatically generated regular expressions. IEEE Access **9**, 38767–38777 (2021)
7. Gibson, A., et al.: Emergency department attendance after telephone triage: a population-based data linkage study. Health Serv. Res. **53**(2), 1137–1162 (2018)
8. Gómez, D., Rojas, A.: An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification. Neural Comput. **28**(1), 216–228 (2016)
9. Kaakinen, P., Kyngäs, H., Tarkiainen, K., Kääriäinen, M.: The effects of intervention on quality of telephone triage at an emergency unit in Finland: nurses' perspective. Int. Emerg. Nurs. **26**, 26–31 (2016)

10. Kadhim, A.I.: Survey on supervised machine learning techniques for automatic text classification. Artif. Intell. Rev. **52**(1), 273–292 (2019). https://doi.org/10.1007/s10462-018-09677-1
11. Kavuluru, R., Rios, A., Lu, Y.: An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. Artif. Intell. Med. **65**(2), 155–166 (2015)
12. Lake, R., et al.: The quality, safety and governance of telephone triage and advice services-an overview of evidence from systematic reviews. BMC Health Serv. Res. **17**(1), 1–10 (2017). https://doi.org/10.1186/s12913-017-2564-x
13. Li, Y., Yang, T.: Word Embedding for Understanding Natural Language: A Survey, vol. 26 (2017). https://doi.org/10.1007/978-3-319-53817-4
14. Marafino, B.J., Boscardin, W.J., Dudley, R.A.: Efficient and sparse feature selection for biomedical text classification via the elastic net: application to ICU risk stratification from nursing notes. J. Biomed. Inf. **54**, 114–120 (2015)
15. Mascio, A., et al.: Comparative analysis of text classification approaches in electronic health records. arXiv preprint arXiv:2005.06624 (2020)
16. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)
17. Mujtaba, G., et al.: Clinical text classification research trends: systematic literature review and open issues. Expert Syst. Appl. **116**, 494–520 (2019)
18. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., Eisenstein, J.: Explainable prediction of medical codes from clinical text. arXiv preprint arXiv:1802.05695 (2018)
19. Mustafa, A., Rahimi Azghadi, M.: Automated machine learning for healthcare and clinical notes analysis. Computers **10**(2), 24 (2021)
20. Navratil-Strawn, J.L., Ozminkowski, R.J., Hartley, S.K.: An economic analysis of a nurse-led telephone triage service. J. Telemedicine Telecare **20**(6), 330–338 (2014)
21. North, F., et al.: Clinical decision support improves quality of telephone triage documentation-an analysis of triage documentation before and after computerized clinical decision support. BMC Med. Inf. Decis. Making **14**(1), 1–10 (2014)
22. Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., Vieira, R.: Assessing the impact of contextual embeddings for Portuguese named entity recognition. In: Proceedings of the 8th Brazilian Conference on Intelligent Systems, pp. 437–442 (2019)
23. Santos, J., dos Santos, H.D.P., Vieira, R.: Fall detection in clinical notes using language models and token classifier. In: Proceedings of the 33rd International Symposium on Computer-Based Medical Systems, CBMS 2020, Rochester, MN, USA, 28–30 July 2020, pp. 283–288 (2020)
24. Santos, J., Terra, J., Consoli, B.S., Vieira, R.: Multidomain contextual embeddings for named entity recognition. In: Proceedings of the 35th Conference of the Spanish Society for Natural Language Processing, pp. 434–441 (2019)
25. Shao, Y., Taylor, S., Marshall, N., Morioka, C., Zeng-Treitler, Q.: Clinical text classification with word embedding features vs. bag-of-words features. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 2874–2878. IEEE (2018)
26. Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P.: Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J. Biomed. Health Inf. **22**(5), 1589–1604 (2017)
27. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds.) BRACIS 2020. LNCS (LNAI), vol. 12319, pp. 403–417. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61377-8_28

28. Stein, R.A., Jaques, P.A., Valiati, J.F.: An analysis of hierarchical text classification using word embeddings. Inf. Sci. **471**, 216–232 (2019)
29. Topaz, M., Woo, K., Ryvicker, M., Zolnoori, M., Cato, K.: Home healthcare clinical notes predict patient hospitalization and emergency department visits. Nursing Res. **69**(6), 448–454 (2020)
30. Tran, D.T., et al.: Compliance with telephone triage advice among adults aged 45 years and older: an Australian data linkage study. BMC Health Serv. Res. **17**(1), 1–13 (2017). https://doi.org/10.1186/s12913-017-2458-y
31. Wagner Filho, J.A., Wilkens, R., Idiart, M., Villavicencio, A.: The brWaC corpus: a new open resource for Brazilian Portuguese. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018). https://www.aclweb.org/anthology/L18-1686
32. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. **13**(3), 55–75 (2018)