Revista Facultad de Ingeniería (Rev. Fac. Ing.) Vol. 33, No. 68.
L-ISSN: 0121-1129, e-ISSN: 2357-5328.

DOI: 10.19053/01211129.v33.n68.2024.18080

AUTOMATIC CLASSIFICATION OF CANCER PATHOLOGY REPORTS WRITTEN IN SPANISH: A MACHINE-LEARNING APPROACH

Clasificación automática de informes de patología del cáncer escritos en español: un enfoque de aprendizaje automático

Nelson-Alejandro Portilla
Universidad del Valle (Cali, Colombia).

Oswaldo Solarte-Pabón
Universidad del Valle (Cali, Colombia).
oswaldo.solarte@correounivalle.edu.co

Luis-Eduardo Bravo
Universidad del Valle (Cali, Colombia).

Fecha de recibido: 01-02-2024
Fecha de aceptado: 30-06-2024

ABSTRACT

Pathology reports represent a rich information resource that describes cancer characteristics in a document written in a free text format. Extracting information from pathology reports is crucial to support cancer research. However, this is a challenging task because these reports are written in medical natural language. In this paper, we propose a machine learning-based approach to extract information automatically from pathology reports written in Spanish. The approach consists of two models: the first one classifies a pathology report into two categories (positive or negative for cancer). The second extracts the cancer tumor site from the pathology text documents. The proposed approach obtains an F1-score of 88% for the first model and 84% for the second model. The obtained results suggest that the proposed approach is feasible to support automatic information from pathology reports written in Spanish.

Keywords: automatic classification; cancer registries; machine-learning; pathology reports.

RESUMEN

Los informes de patología representan una rica fuente de información que describe las características del cáncer en un documento escrito en formato de texto libre. Extraer información de los informes de patología es crucial para respaldar la investigación del cáncer. Sin embargo, esta es una tarea desafiante porque los informes están escritos en lenguaje natural médico. En este artículo, proponemos un enfoque basado en aprendizaje automático para extraer información automáticamente de informes de patología escritos en español. El enfoque consta de dos modelos: el primero clasifica un informe de patología en dos categorías (positivo o negativo para cáncer). El segundo extrae la ubicación del tumor canceroso de los documentos de texto de patología. El enfoque propuesto obtiene un puntaje F1 del 88% para el primer modelo y del 84% para el segundo modelo. Los resultados obtenidos sugieren que el enfoque propuesto es factible para soportar la extracción automática de los informes de patología escritos en español.

Palabras clave: aprendizaje automático; clasificación automática; registros de cáncer; reportes de patología.

How to cite: N. A. Portilla, O. Solarte Pabón, and L. E. Bravo, Automatic Classification of Cancer Pathology Reports Written in Spanish: A Machine-Learning Approach. *Revista Facultad de Ingeniería*, vol. 33, no. 68, e18080, 2024. https://doi.org/10.19053/01211129.v33.n68.2024.18080

1. INTRODUCTION

Cancer registries process high volumes of free-text reports, a valuable source of information regarding cancer diagnosis [1], [2]. Extracting information from pathology reports is useful to support clinical research and improve patient care [6], [7], [8]; these reports are written in medical natural language [9] by pathologists who are experts in interpreting laboratory tests. They study samples and tissues from a patient and write the results in a free text form [3], [4]. Reports represent rich information that describes cancer characteristics in detail and are commonly processed in cancer registries for human-trained coders who read, interpret, and manually extract information such as the cancer diagnosis, tumor location in the body, tumor grade, etc. In addition, cancer registers must be validated and integrated among hospitals or local healthcare resource curated information about pathology results analysis [5]. Information extraction and coding cancer diagnosis from textual data sources have been performed using a manual labor-intensive process, which is significantly expensive and time-consuming. To deal with this problem researchers have proposed rule-based [10]-[11], machine learning-based [12], [13], [14], or deep learning-based solutions [15], [16], [17].

Moreover, most of the proposals described above assume that extracted information is written affirmatively, but it is known that negation is widely used in medical texts, and detecting it correctly improves the extraction of information. Therefore, a component to detect negation should be present in proposals aimed at extracting information from pathology reports. This is crucial to correctly extract medical diagnoses [18], [19], [20].

Although several studies have proposed interesting approaches for extracting automatic information from cancer pathology reports, most of them focused on the English language [21]. Hence, extracting information from clinical resources written in languages different from English poses some challenges [22]. In the case of Spanish, most efforts have concentrated on extracting named entities or medical concepts from clinical narratives. Automatic classification of cancer pathology reports written in Spanish has not been addressed in depth and efforts are required to improve this process [23].

In this paper, we propose a machine learning-based approach to classify cancer pathology reports written in Spanish automatically. It consists of two steps: the first one classifies a pathology report into two possible categories (positive or negative for cancer); in the second, the approach predicts the primary cancer site in the body from reports classified as positive in the previous step. The main contributions of this paper are a corpus manually annotated by medical experts that contains a set of cancer pathology reports with two annotated categories: Positive and negative for cancer and the primary site of the tumor; and a machine learning-based approach to automatically process pathology reports and classify them. This approach opens opportunities for the medical community to improve medical information analysis in Spanish.

The remainder of this paper is organized as follows: Section 2 describes materials and methods proposed in this approach; Section 3 describes experimentations and results; Section 4 describes the discussion of the results obtained; finally, Section 5 presents conclusions and future work.

2. METHODOLOGY

In this section, we will present the approach to the automatic classification of pathology reports written in Spanish.

A. Dataset

The dataset used in this study consists of 10,299 pathology reports from different laboratories, which were entered in the population cancer registry from Cali, Colombia. These pathology reports were manually annotated by four expert pathologists. Each pathology report in the dataset contains three columns as shown in Table 1, described as follows:

- **Pathology result:** it describes the diagnosis reported by the laboratory after a piece of tissue is examined by a pathologist. This diagnosis description is written in a narrative form using medical natural language (Table contains the description in Spanish).
- **Cancer diagnosis status:** it is a manual annotation performed by expert human annotators in the cancer registry. The diagnosis status annotation can take two values: cancer positive (1) and cancer negative (0). The dataset contains 3399 positive reports and 6900 negative reports.
- **Primary tumor site:** it is also a manual annotation that indicates the primary site of the cancer tumor. The cancer registry has annotated thirteen different primary tumor sites. Table 2 shows different cancer tumor sites in the dataset.

Table 1. Pathology reports dataset used in this study (In Spanish)

Pathology Result	Cancer Status	Tumor Site
Glándula mamaria izquierda lesión biopsia estudios que confirman antecedente de carcinoma mamario- español	1	Cáncer de mama
Sintomática respiratoria más masa en pulmón. Sin evidencia de proceso neoplásico maligno en la muestra tejido adyacente con edema hemorragia y congestión.	0	Cáncer de pulmón

Table 2. Number of tumor sites for primary tumors evaluated by the experts from the Cancer Registry

Tumor site	Number of examples	Tumor site	Number of examples
Breast organ	645	Lymph nodes	170
Brain	599	Prostate	157
Thyroid	282	Bronchial and Lung	139
Stomach	174	Prostate	157
Colon	106	Cervix	83
Ovary	83	Kidney	80

B. Feature Extraction

The feature extraction process is shown in Figure 1. It is divided into three steps: preprocessing, cancer concept annotation, and negation detection. The process of each component will be described below. Features will be explained at the end of this section.

Preprocessing: in this step, pathology reports are pre-processed using natural language processing (NLP) approaches. These include text splitting, tokenization, accent removal, text cleaning, and sentence detection.

Cancer concept annotation: in this step, we combine the UMLS (Unified Medical Language System) database, medical dictionaries created by medical experts, and rules to extract cancer information. We used regular expressions to match information in the UMLS database to information in the pathology

report. This approach uses natural language processing to identify named entities within clinical texts written in Spanish through regular expressions and find specific patterns in the text.

We analyze each term described in the medical dictionary and search for it in the UMLS database. If the search produces results associated with the category *"Neoplastic Process"* in the UMLS database, then, we associate the analyzed report with the cancer category. The *"Neoplastic Process"* category is associated with cancer in UMLS. Table 3 shows a set of features related to concept annotation used in this approach. We used twelve features for training models that extract the cancer status.

Negation Detection

The third step consist of identifying negative concepts in the pathology report. Detecting negation is crucial to processing medical text because it indicates if a medical concept is affirmative or negative. To detect negation, we use the approach described in [20], where each token in a medical text is classified into negative or affirmative. The approach proposed in [10] addresses the negation detection into two steps: cue detection and scope identification. The cue is the word indicating negation, the scope is the set of words affected by a negation cue.

For example, in the sentence: *"Masa en mama izquierda,* ***negativo*** *para carcinoma ductal infiltrante"*, the cue is shown in bold, and the scope is underlined. In this example, the diagnosis of *"carcinoma ductal infiltrante"* is affected by the negation cue, therefore, the concept is negative for cancer.

Negation detection aims to identify words in a pathology report that the physician have negated and indicates a negative status for a possible cancer diagnosis. It is a complex problem, and it is addressed in other studies such as the proposal described in [20].

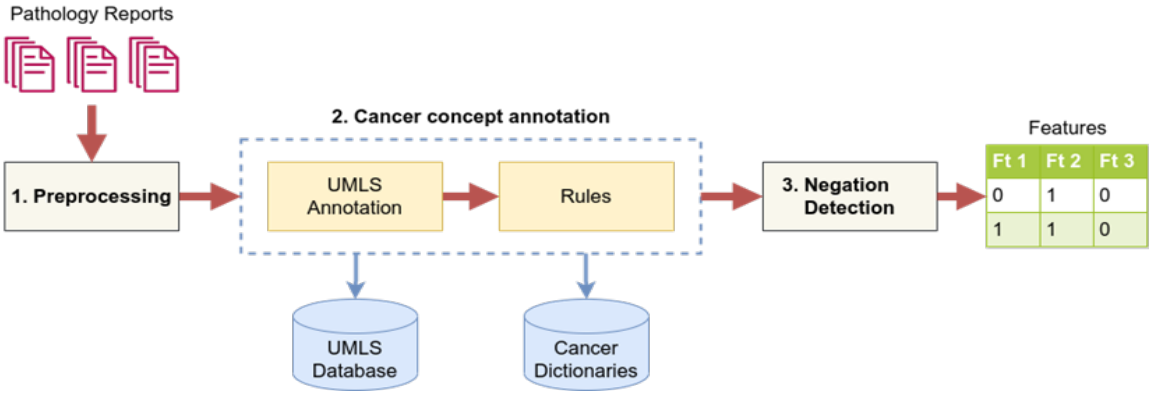

Figure 1. *Feature extraction process*

To extract the primary site of the cancer, we use the TD-IDF model [23]. First, the text is cleaned, accents and special characters are removed; then, the text is divided into words or tokens; finally, a *NxM* matrix is built, where *N* corresponds to the number of pathology reports, and *M* to the number of text features. In this case, the features in the matrix refer to the words contained in the pathology reports that represent the category of organ or neoplasm in the UMLS database.

Table 3. Set of features used in the proposed approach

Feature	Description
1	Evaluate if the pathology report contains at least one concept categorized as "Neoplastic Process" in UMLS.
2	Evaluate if the pathology report contains at least one concept categorized as "Neoplastic Process" in UMLS and that it is not negated.
3	Evaluate if the pathology report contains all the concepts categorized as "Neoplastic Process" in UMLS and that are negated.
4	Evaluate if the pathology report has an instance in the table "annotations."
5	Evaluate if the pathology report contains at least one concept categorized as "Neoplastic Process" in UMLS and is found in the dictionary of benign neoplasms.
6	Evaluate if the pathology report contains all the concepts categorized as "Neoplastic Process" in UMLS is not negated, and is found in the dictionary of benign neoplasms.
7	Evaluate if the pathology report has at least one concept categorized as "Receptor" and is equal to "mom" or "mammary gland."
8	Evaluate if the pathology report does not have an instance in the table "concepts."
9	Evaluate if the pathology report contains at least one concept categorized as "Neoplastic Process" in UMLS and that they are denied.

C. Machine Learning Algorithms

In this section, we describe the machine learning algorithms used to generate the models for predicting the cancer status and the tumor site using the dataset described in section 3.1. Decision trees, Naive Bayes, Multilayer Perception (MLP), Nearest Centroid Classifier, and Stochastic Gradient Descent were used and a set of hyperparameters was configured for each algorithm, as described in Table 4:

- **Decision trees:** depth ("max_depth") determines the maximum depth of tree generation. The minimum number of samples ("min_samples_split") consists of the minimum number of samples needed in a leaf node. The random state ("random_state") controls the randomness of the estimator.
- **Support Vector Machines (SVM)**: kernel defines the methods used by the algorithm to take input data and transform it to the specified type, in this case, to the linear type among the other options (poly, RBF, Sigmoid, and Precomputed). The shrinking parameter helps to shorten the training time if the number of iterations is large. The "probability" parameter helps define the probability estimates of class membership. The parameter "gamma" is the kernel coefficient and defines how far the influence of a single training example reaches, with low values meaning "far" and high values meaning "near", for this case, the scalable value is defined as 1 / (n_features * X.var()). The other possible options are auto, which is defined as 1/n_features. The "class_weight" parameter defines the weight of the classes, in this case, all classes weigh the same and have a value equal to one.
- **Multi-layer Perceptron (MCP)**: the "Hidden_layer_size" defines the size and number of neurons of the hidden layers, the size of the tuple shows the number of hidden layers and each item of the tuple refers to the number of neurons in each layer. In this case, three layers are used: two correspond to the input and output, one corresponds to the hidden layer and 4 neurons for the hidden layer. 'Solver' defines the method for weight optimization, in this case, 'lbfgs' is an optimizer from the quasi-Newton family of methods, and cross-validation determines it as the best between the SGD and Adam options. Learning rate is defined as constant (invscaling and adaptive are other options). For the activation function, options are identity, logistic, tanh, and Relu; the logistic

activation function is the one with the best precision according to the GridSearchCV method. The 'alpha' parameter defines the penalty coefficient L2 or regulation term and the numerical value 0.00001 is established.

- **Stochastic Gradient Descent (SGD)**: the implementation uses the SGDClassifier class, which implements a simple stochastic gradient descent learning routine that supports different loss functions and penalties for classification. Like other classifiers, SGD has to be equipped with two matrices: an X matrix that corresponds to the features matrix, and a vector that contains the values of the evaluated class. Table 4 shows the assigned parameters. The validation method is configured with the hinge, log, modified_huber, and squared_hinge options for the loss function 'loss', which defines the hinge method as the best option; then, a constant learning rate is defined between the options (optimal, inverse scaling, and adaptive) and the coefficient 'eta0' that defines the initial learning rate is determined to be 0.1 within the possible range (zero to one).

Table 4. Hyperparameters used to configure algorithms

Algorithm	Hyperparameters
Decision Tree	Max_Depth = 10 Min_Samples_Split = 2 Random_state = 1
Support Vector Machines (SVM)	Kernel = Linear Shrinking = True Probability= True
Multilayer Perceptron	Alpha =1e-05 activation = logistics learning_rate= constant hidden_layer = 4
Stochastic Gradient Descent	Loss = Hinge learning_rate =constant eta0 = 0.1

3. EXPERIMENTATION AND RESULTS

For training the algorithms described above, we used a cross-validation strategy with $k = 10$. Moreover, the method *"GridSearchCV"* was used to find the hyperparameters shown in Table 4. The Python library *Scikit-learn* provides a method for using the Grid search over multiple hyperparameters. The metrics for the evaluation of the models are standard metrics such as Precision, Recall, and F1-score. Results were calculated as the average of all iterations in the cross-validation strategy.

Table 5 shows the results for the cancer status classification task. This is a binary classification problem that categorizes a pathology report with a positive or negative label. According to Table 5, the proposed approach obtains feasible results. Specifically, the Decision tree and the Multi-layer Perceptron have obtained the best performance. Decision tree obtained an F1-score of 89% while the Multi-layer Perceptron obtained 88%. Moreover, the decision tree algorithm contributed the fewest false negatives, it has the highest sensitivity value concerning the other models. In the cancer domain, sensitivity is crucial, and it is the one that should be minimized as much as possible since they are patients that the model predicts as negative, but they are really cancer patients.

In the medical field, it is very important to take into account all patients, and to avoid falling into underestimation of statistics, the reading of 100% of the cases should be approximated. Specifically, in the Cancer Registry, two concepts are taken into account: exhaustiveness and timeliness. The first

consists precisely of collecting the greatest number of cases from all the institutions; the opportunity is to deliver results such as reports and statistics as quickly and promptly as possible.

Table 5. Results for cancer status classification (Model 1)

Model	Precision	Recall	F1-score
Decision Tree	**0.93**	**0.85**	**0.88**
SVM	0.91	0.80	0.85
MCP	0.93	0.83	0.87
SGD	0.89	0.80	0.83

The proposed models show an F1-score above 80% and several false negatives between 61 and 68 cases out of 353 total with cancer, that is, a proportion of 17%. Therefore, the exhaustiveness would be affected in that proportion (17%), considering that the false positives together with the true positives are later reviewed by experts, and they can be discarded in the process.

Automatic extraction processes and classification models presented above make them relevant tools for improving the timeliness of information. However, to close the gap in terms of completeness, the performance of the models must continue to be improved to obtain a lower number of false negatives.

Tables 6 and 7 show the results obtained for the second task addressed in this paper: classifying the tumor site in the body from a pathology report. Table 6 describes the Decision tree performance while Table 7 describes that of the Multi-layer perception (MCP) approach. These results show that the MCP-based approach outperforms the Decision tree-based approach. Results obtained from the MCP show an F1-score above 80%, which means that this approach is capable of extracting the cancer tumor site from pathology reports written in Spanish. The best performance was obtained for "Thyroid cancer" (Cáncer de tiroides) and "Breast cancer" (Cáncer de mama) obtaining an F1-score of 92% and 94%, respectively.

Table 6. Results on extracting tumor site (Decision Tree)

Tumor Site	Precision	Recall	F1-score
Bronchial	0.71	0.43	0.53
Colon	0.67	0.48	0.56
Cervix	0.73	0.47	0.57
Brain	0.84	0.38	0.53
Stomach	0.63	0.68	0.66
Lymph node	0.67	0.61	0.63
Thyroid	0.79	0.71	0.75
Breast organ	0.85	0.85	0.85
Ovary	0.64	0.38	0.47
Prostate	0.88	0.71	0.79
Average	**0.77**	**0.58**	**0.63**

Decision tree-based approach performance is lower compared to the MCP-based approach. This can be explained by the small number of examples used as support to classify the tumor site, which suggests the opportunity to increase the number of examples in the dataset.

Table 7. Results on extracting tumor site (Multilayer Perceptron)

Tumor site	Precision	Recall	F1-score
Bronchial	0.85	0.94	0.89
Colon	0.88	0.82	0.85
Cervix	0.90	0.82	0.86
Brain	0.84	0.90	0.84
Stomach	0.91	0.83	0.87
Lymph node	0.68	0.61	0.64
Thyroid	0.93	0.90	0.92
Breast Organ	0.92	0.96	0.94
Ovary	0.82	0.88	0.84
Prostate	0.94	0.91	0.93
Average	**0.87**	**0.82**	**0.84**

4. DISCUSSION

The evaluation of 32,000 pathology reports by a team of experts from the Cancer Registry lasted six months, while the automatic information extraction process took 48 hours. Therefore, this method improves the time and cost of analyzing information from cancer registries. Resource optimization is essential in low- to middle-income countries like Colombia.

Decision Tree and Multi-layer perception (MCP) models performed better because these algorithms have fault tolerance. The accuracy of MCP could improve with increased computational power.

The MCP model was selected to conduct the evaluation and adjustment tests for the classification process. The sensitivity of MCP was 11% higher than the conventional classification method used by the Cancer Registry, 92% vs. 81%. The addition of two features and the exclusion of flow cytometry pathology reports increased the sensitivity of the ANN model to 92% and produced no change in specificity. Although the specificity was higher with the conventional method (94% vs 90%), false 312 positive cases can be detected and excluded by the Cancer Registry during the validation and coding process of cancer cases. The MCP model is a reliable source for the automatic classification process of cancer diagnoses obtained from pathology reports written in Spanish. However, it is necessary to continue tuning the model to improve sensitivity.

5. CONCLUSIONS

The most important source of information to extract cases from a cancer registry is the pathology report because it contains precise diagnostic data obtained through the morphological study of cells and tissues. However, a very high proportion of cases diagnosed by histology or cytology/hematology suggests incompleteness of the cancer registry due to excessive reliance on the pathology laboratory as a source of information, with failure to find cases diagnosed by other means (clinical diagnosis, surgical and imaging diagnosis, death certificate as the only evidence of cancer). As future work, we are planning to increase the number of instances in the dataset and test deep learning approaches to extract information from pathology reports.

AUTHORS' CONTRIBUTIONS

Oswaldo Solarte: methodology, research, writing-original draft.

Nelson Portilla: methodology, software, writing-review and editing.

Luis Eduardo Bravo: validation, reviewed the experiments, writing-review and editing.

REFERENCES

[1] K. Liu, KJ. Mitchell, W. Chapman. "Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set. AMIA," in *Annual Symposium Proceedings*, 2005, pp. 460-464.

[2] V. Jouhet, V. Defossez, G. Burgun, A. Le Beux, P. Levillain, P. Ingrand, P. Claveau, "Automated classification of free-text pathology reports for registration of incident cases of cancer," *Methods of Information in Medicine*, vol. 51, pp. 242-251, 2012. https://doi.org/10.3414/ME11-01-0005

[3] E. Soysal, J. L. Warner, J. Wang, K. Harvey. "Developing customizable cancer information extraction modules for pathology reports using clamp," *Studies in Health Technology and Informatics*, vol. 264, pp. 1041-1045, 2019. https://doi.org/10.3233/SHTI190383

[4] R. Kavuluru, I. Hands, E. B. Durbin, L. Witt, "Automatic Extraction of ICD-O-3 Primary Sites from Cancer Pathology Reports," in *AMIA Proceedings*, 2013, pp. 112-116.

[5] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum. "Clinical Natural Language Processing in languages other than English: Opportunities and challenges," *Journal of Biomedical Semantics*, vol. 9, pp. 1-13, 2018. https://doi.org/10.1186/s13326-018-0179-8

[6] P. López-Úbeda, T. Martín-Noguerol, J. Aneiros-Fernández, A. Luna, "Natural Language Processing in Pathology: Current Trends and Future Insights," *American Journal of Pathology*, vol. 192, no. 11, pp. 1486-1495, 2022.

[7] M. Sarango, R. Reátegui, "Medical Entities Extraction with Metamap and cTAKES from Spanish Texts," in *International Conference on Information Technology & Systems*, 2023, pp. 197-203.

[8] K. G Zeng, H. Zhang, M. A. Harbi, "Improving Information Extraction from Pathology Reports using Named Entity Recognition," *Research Square*, Preprint, 2023. https://doi.org/10.21203/rs.3.rs-3035772/v1

[9] H. Dalianis. "Clinical Text Mining", in *Springer Open*, 2018. https://doi.org/10.1007/978-3-319-78503-5

[10] R. S. Crowley, M. Castine, K. Mitchell, G. Chavan. "CATIES: A grid-based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research," *Journal of the American Medical Informatics Association*, vol. 17, pp. 253-264, 2010. https://doi.org/10.1136/jamia.2009.002295

[11] A. Nguyen, J. Moore, G. Zuccon, M. Lawley, S. Colquist, "Classification of pathology reports for cancer registry notifications," in *Health Informatics: Building a Healthcare Future Through Trusted Information*, 2012, pp. 150-156.

[12] S. Martina, L. Ventura, P. Frasconi, "Classification of cancer pathology reports: A large-scale comparative study," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 3085-3094, 2020. https://doi.org/10.1109/JBHI.2020.3005016

[13] T. Oliwa, S. B Maron, L. M. Chase, S. Lomnicki, D. V. Catenacci, "Obtaining Knowledge in Pathology Reports Through a Natural Language Processing Approach with Classification, Named-Entity Recognition, and Relation-Extraction Heuristics," *Journal of Clinical Cancer Informatics*, vol. 8, pp. 4051-4058, 2019. https://doi.org/10.1200/cci.19.00008.406

[14] G. Napolitano, A. Marshall, P. Hamilton, A. Gavin, "Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction," *Artificial Intelligence in Medicine*, vol. 70, pp. 77-83, 2016. https://doi.org/10.1016/j.artmed.2016.06.001

[15] T. Mikolov, I. Sutskever, K. Chen, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111-3119. https://doi.org/10.48550/arXiv.1310.4546

[16] J. Pennington, R. Socher, C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Empirical Methods in Natural Language Processing*, 2014, pp. 1532-1543. https://doi.org/10.3115/v1/D14-1162

[17] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135-146, 2017. https://doi.org/10.1162/tacl_a_00051

[18] J. X Qiu, H. J Yoon, P. A Fearn, G. D, Tourassi, "Deep Learning for Automated Extraction of Primary Sites from Cancer Pathology Reports," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, pp. 244-251, 2018. https://doi.org/10.1109/JBHI.2017.2700722

[19] S. Martina, L. Ventura, P. Frasconi, "Classification of cancer pathology reports: A large-scale comparative study," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 3085-3094, 2020. https://doi.org/10.1109/JBHI.2020.3005016

[20] O. Pabón, O. Montenegro, M. Torrente, A. R. González, M. Provencio, E. Menasalvas, "Negation and uncertainty detection in clinical texts written in Spanish: a deep learning-based approach," *PeerJ Computer Science*, vol. 8, e913, 2022. https://doi.org/10.7717/peerj-cs.913

[21] A. Khandelwal, S. Sawant, "NegBERT: A transfer learning approach for negation detection and scope resolution," in *12th International Conference on LanguageResources and Evaluation, Conference Proceedings*, 2020, pp. 5739-5748. https://doi.org/10.48550/arXiv.1911.0421

[22] O. Solarte-Pabón, A. Blazquez-Herranz, M. Torrente, A. Rodríguez-Gonzalez, M. Provencio, E. Menasalvas, "Extracting Cancer Treatments from Clinical Text written in Spanish: A Deep Learning Approach," in *8th International Conference on Data Science and Advanced Analytics*, 2021, pp. 1-6. https://doi.org/10.1109/DSAA53316.2021.9564137

[23] C. Condordan, M. Cheryl. C. Benavides, G.Cecchi. "Natural language processing: Oportunnities and challenges for patients, providers, and hospital systems," *Psychiatric Annals*, vol. 49, no. 5, pp. 202-208, 2019. https://doi.org/10.3928/00485713-20190411-01

redalyc
UAEMEX

Available in:
https://www.redalyc.org/articulo.oa?id=413982266005

How to cite

Complete issue

More information about this article

Journal's webpage in redalyc.org

Scientific Information System Redalyc
Diamond Open Access scientific journal network
Non-commercial open infrastructure owned by academia

Nelson-Alejandro Portilla, Oswaldo Solarte-Pabón,
Luis-Eduardo Bravo

Automatic classification of cancer pathology reports written in spanish: a machine-learning approach
Clasificación automática de informes de patología del cáncer escritos en español: un enfoque de aprendizaje automático

Revista Facultad de Ingeniería
vol. 33, no. 68, e18080, 2024
Universidad Pedagógica y Tecnológica de Colombia,
ISSN: 0121-1129
ISSN-E: 2357-5328

DOI: https://doi.org/10.19053/01211129.v33.n68.2024.18080