Técnicas de Clasificación de Documentos de Literatura Médica Empleando Active Learning y Word Embedding

Classification Techniques for Medical Literature Documents Using Active Learning and Word Embedding

Benjamín Ayancán Guerrero[1], alumno de 5º año.
Denis Parra Santander[1], profesor asociado.

[1]Departamento de Ciencia de la Computación, Escuela de Ingeniería, Pontificia Universidad Católica de Chile.

*Autor para correspondencia: dparra@ing.puc.cl

Benjamín Ayancán Guerrero[1], 5th year student.
Denis Parra Santander[2], associate professor.

[1]Department of Computer Science, School of Engineering, Pontificia Universidad Católica de Chile.

*Corresponding author: dparra@ing.puc.cl

RESUMEN

En las revisiones sistemáticas de literatura médica, el creciente número de estudios publicados implica un trabajo de selección para los revisores, quienes pueden llegar a examinar miles de artículos depositados en bases de datos y sistemas de indexación. Trabajos anteriores han modificado la codificación de textos para mejorar su representación, sin embargo, estos enfoques no ahondan en la desproporcionalidad de clases en *data sets* con deficiencias de construcción.

En este contexto, *Active Learning* (AL) permite seleccionar aquellos datos más relevantes para etiquetar, reduciendo tanto la cantidad requerida como el costo asociado. En este trabajo, evaluamos la incidencia de modelos de lenguaje neuronal BERT y Word2Vec, además del entrenamiento con AL y *Data Augmentation* (DA) para manejar la asimetría en los data sets incluidos en el desafío CLEF eHealth 2017.

Las experimentaciones arrojaron un impacto positivo en el tratamiento del problema. Para clasificación tradicional, Random Forest con BERT y DA logra un 0,861 de AUC y un f1-score de 0,83 en la clase relevante. Además, Word2Vec pasa de un f1-score de 0,08 a un 0,31 sobre las respuestas relevantes y AUC de 0,591 usando solamente el 2.6% de los datos totales en el entrenamiento.

Palabras clave: active learning, aprendizaje automático, literatura médica, natural language processing, word embedding.

ABSTRACT

In systematic reviews of the medical literature, the growing number of studies demands extensive screening from reviewers, who manually examine thousands of articles. In previous studies, variations in text encoding have been explored looking for greater representation efficiency. However, these approaches have not delved into the disproportionality of classes contained in data sets due to deficiencies in their construction.

In this context, Active Learning allows to select the most relevant data for labeling, reducing both the required amount and the cost of obtaining them. We focused on evaluating the incidence of neural language models such as BERT and Word2Vec, training with Active Learning, and proposing a data augmentation method to tackle symmetry problems in the CLEF eHealth 2017 data set.

Our research showed that the proposed techniques positively impacts class imbalance. For traditional supervised classification, Random Forest with BERT Embedding combined with our method achieves a 0.861 AUC, along with an f1-score of 0.83 in the relevant class. On the other hand, Word2Vec goes from an f1-score of 0.08 to 0.31 on relevant responses, reaching an AUC of 0.591 using only 2.6% of the total training data.

Keywords: active learning, automatic learning, medical literature, natural language processing, word embedding.

1. INTRODUCCIÓN

1.1 Contexto y Antecedentes

El proceso de revisión sistemática de artículos para reunir información de manera fidedigna es esencial en medicina, donde la selección de referencias es de vital importancia (Gurevitch et al., 2018). En este contexto, Sackett et al. (1996) definen la Medicina Basada en Evidencia (EBM) como el uso consciente, explícito y juicioso de la mejor evidencia actual para tomar decisiones sobre la atención de pacientes individuales. Esta es una práctica recurrente, aunque ineficiente sin una técnica adecuada (Maynard, 1997), ya que la evidencia se obtiene de información distribuida en diversas fuentes (Canese & Weis, 2013; Doms & Schroeder, 2005).

1. INTRODUCTION

1.1 Context and Background

The process of systematic review of articles to gather reliable information is essential in medicine, where the selection of references is of vital importance (Gurevitch et al., 2018). In this context, Sackett et al. (1996) define Evidence-Based Medicine (EBM) as the conscious, explicit and judicious use of the best current evidence to make decisions about the care of individual patients. This is a recurring practice, although inefficient without an adequate technique (Maynard, 1997) because the evidence is obtained from information distributed in various sources (Canese et al., 2013; Doms et al., 2005).

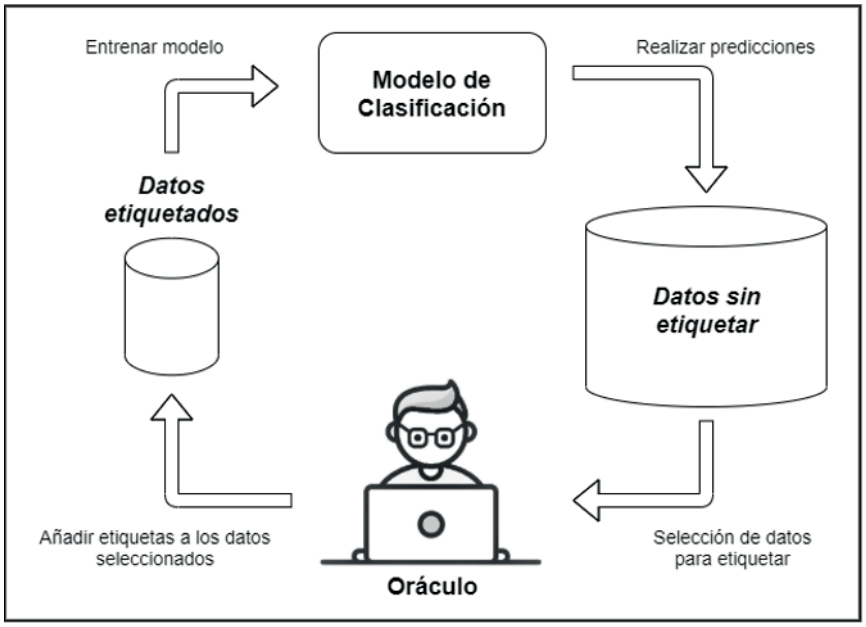

Figura 1. Ilustración del método de *Active Learning* (AL). A partir de un conjunto de datos sin etiquetar, estos son seleccionados a través de una metodología para su revisión. El oráculo entrega una etiqueta a los datos de forma correcta y son añadidos al conjunto. Se reentrena el modelo con los datos etiquetados para luego realizar predicciones y evaluaciones del método.

Figure 1. Illustration of the Active Learning (AL) approach. From a set of unlabeled data, these are selected using a review methodology. The oracle assigns a correct label to the data and they are added to the ready-made set. The model is retrained with these labeled data and then can make predictions and evaluations of the method.

El etiquetado de los datos representa un cuello de botella en el aprendizaje automático, especialmente en *natural language processing* (NLP), donde el costo es bastante abultado (Strubell et al., 2019). Las metodologías que involucran *Active Learning* (AL) proponen que el entrenamiento de un modelo de clasificación sea más económico (Settles, 2009; Baldridge & Osborne, 2004), pues bajo esta fórmula el propio algoritmo escoge los mejores datos para entrenar, necesitando así una menor cantidad (Settles, 2011). Como se muestra en la Figura 1, el algoritmo busca reconocer de forma iterativa los datos más relevantes y luego consultar las etiquetas a un **ORÁCULO**.

Esto ayuda en áreas con deficiencias en la generación de sus conjuntos de datos, por ejemplo, en textos médicos debido a problemas de privacidad y/o escasez de experiencia (Dernoncourt et al., 2017). Las implementaciones de las últimas décadas han mostrado que AL es efectivo en la reducción del tamaño del conjunto de entrenamiento, acelerando el proceso de aprendizaje con una fracción de la participación humana (Settles, 2011).

1.2 Trabajo Relacionado

Carvallo et al. (2020) realizaron una comparación de los **EMBEDDINGS** Word2Vec, Glove y BERT para la clasificación con *Active Learning* sobre los conjuntos CLEF

Labeling of data represents a bottleneck in machine learning, especially in natural language processing (NLP), where the cost is quite high (Strubell et al., 2019). Methodologies that involve Active Learning (AL) involve a more economical training of classification models (Settles et al., 2009) (Baldridge et al., 2004), since under this formula the algorithm itself chooses the best training data, thus requiring a reduced amount (Settles et al., 2011). As shown in **Figure 1**, the algorithm seeks to iteratively recognize the most relevant data and then query an **ORACLE** for the labels.

This is helpful in areas with deficiencies in the generation of data sets, for example, in medical texts due to privacy concerns and/or lack of experience (Dernoncourt et al., 2017). Implementations in the last decades have shown that AL is effective in reducing the size of the training set, accelerating the learning process with a fraction of human participation (Settles et al., 2011).

1.2 Related Work

Carvallo et al. (2020) compared the EMBEDDINGS Word2Vec, Glove and BERT for classification with Active Learning on the CLEF and Epistemonikos datasets in order to find the best performing combination.

y Epistemonikos, con el fin de encontrar la combinación con mejor rendimiento. Su trabajo reveló que los modelos BioBert proporcionan un mejor rendimiento (Carvallo et al., 2020). Sin embargo, no aborda una estrategia para contrarrestar la desproporcionalidad que presentaba el dataset CLEF. En este ámbito, concluyen que Epistemonikos presenta una menor complejidad para trabajar.

Por otra parte, Yuan estudia un enfoque auto supervisado con *Active Learning*, para disminuir los costos emplea una estrategia de representación por medio de enmascaramiento (Yuan et al., 2020). A pesar de ello, se limita a usar muestras balanceadas de los conjuntos originales.

En esta investigación se realiza un recorrido de la incidencia de técnicas en la clasificación considerando modelos de lenguaje neuronal BERT (Devlin et al., 2018) y Word2Vec (Mikolov et al., 2013), como también se analiza el uso de *Active Learning*. Adicionalmente, se proponen implementaciones con enmascaramiento para las muestras de incertezas en entrenamiento con *Active Learning*, junto con una técnica de *Data Augmentation* (DA) sobre las consultas con el fin de proporcionar mejor el conjunto (Ribeiro et al., 2018).

2. METODOLOGÍAS DE EXPERIMENTACIÓN

2.1 Dataset

Para los experimentos usamos CLEF eHealth 2017 (Goeuriot et al., 2017). Estos documentos en inglés están compuestos por 49 preguntas de literatura médica con 190.228 posibles artículos respuesta, conformando 434.643 pares de evidencia relevante o no relevante. Se realizó una limpieza preliminar de datos con defectos, resultando en la selección de 260.000 pares consulta-artículo de los cuales el 98,2% posee una etiqueta de no relevante y 1,7% de relevante. Para los experimentos, los textos se concatenaron en consulta, título y resumen del artículo.

2.2 Tokenización con Word2Vec y BERT

Para el primer caso se entrenó Word2Vec con el método CBOW, pasando cada oración a un vector de largo 300 promediando las representaciones de las palabras. Por otro lado, BERT utilizó un transformador pre entrenado para obtener un vector de 512 dimensiones por cada oración.

2.3 *Data Augmentation*

Para manejar el sobreajuste en la clasificación por la desproporcionalidad se propone una técnica para nivelar el porcentaje de pares relevantes. Usando *Fill-Mask* de *Hugging Face*, implementamos un proceso de enmascaramiento aleatorio, donde se predice una palabra con el modelo para reemplazar generando una variante a la consulta original.

Their work revealed that BioBERT models provide better performance (Carvallo et al., 2020). However, they did not devise a strategy to counteract the disproportionality of the CLEF dataset. In this respect, they concluded that Epistemonikos offers less complexity to work with.

On the other hand, Yuan studied a self-supervised approach with Active Learning, using a masking representation strategy to reduce costs (Yuan et al., 2020). Nonetheless, this is still limited to balanced samples from the original sets.

In this study, we carry out a survey of the incidence of techniques in the classification considering the neural language models BERT (Devlin et al., 2018) and Word2Vec (Mikolov et al., 2013), as well as an analysis of the use of Active Learning. Additionally, we propose implementations with masking for the training uncertainty samples with Active Learning, together with a Data Augmentation (DA) technique on the queries in order to improve the proportions of the set (Ribeiro et al., 2018).

2. RESEARCH METHODOLOGY

2.1 Dataset

To run the experiments, we used the CLEF eHealth 2017 dataset (Goeuriot et al., 2017). These documents in English are composed of 49 questions from the medical literature with 190,228 possible response articles, making up 434,643 pairs of relevant or not relevant evidence. We performed a preliminary cleaning of defective data, resulting in the selection of 260,000 query-article pairs, of which 98.2% have a label of not relevant and 1.7% are labeled relevant. In the experiments, texts were concatenated in query, title and abstract of the article.

2.2 Tokenization with Word2Vec and BERT

In the first case, Word2Vec was trained using the CBOW method, passing each sentence to a vector of length 300, averaging the representations of the words. On the other hand, BERT used a pre-trained transformer to obtain a 512-dimensional vector for each sentence.

2.3 Data Augmentation

To manage the overfitting in the classification due to disproportionality, we propose a technique that levels the percentage of relevant pairs. Using Hugging Face's Fill-Mask, we implemented a random masking process, where a word is predicted with the model and it is replaced, generating a variant to the original query. This technique only applies to queries, since the articles in the medical literature are external sources.

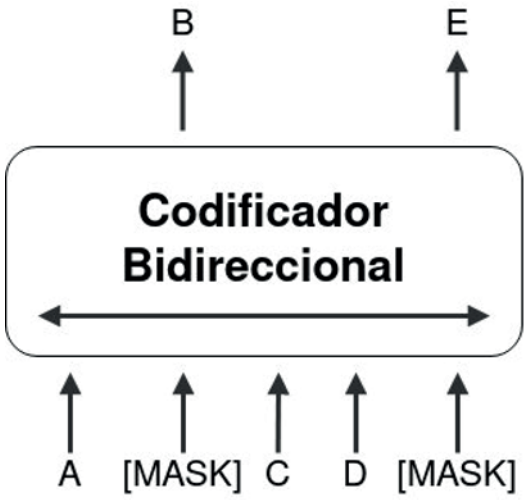

Figura 2. Tokens aleatorios son reemplazan por una máscara, denotada por [MASK], para que el modelo DistilBERT prediga de forma individual cada palabra (B, E) a partir del contexto de ambos lados (A, C, D) (Lewis et al., 2019) (Wolf et al., 2019).

Figure 2. Random tokens are replaced by a mask, denoted by [MASK], so that the DistilBERT model individually predicts each word (B, E) through the context of both sides (A, C, D) (Lewis et al., 2019) (Wolf et al., 2019).

Esta técnica solo se aplica sobre las consultas, debido a que los artículos de la literatura médica son fuentes externas.

2.4 Representación por enmascaramiento

El modelado de lenguaje enmascarado es la tarea de enmascarar **TOKENS** en una secuencia y pedirle al método que llene esa máscara con un token apropiado. Para completar la actividad se utiliza el modelo BERT de representaciones de codificaciones con *transformers* (Devlin et al., 2018).

Yuan propone, a través del enmascaramiento, potenciar la selección por incertidumbre en el modelo de *Active Learning* (Yuan et al., 2020). Basándose en este enfoque, se aplica un enmascaramiento aleatorio en las oraciones con *Fill-Mask*, representando de forma más ligera.

2.5 Clasificación Supervisada Tradicional y Active Learning

Para este experimento nos enfocándonos en dos métodos: *Random Forest* y Regresión Logística (Carvallo et al., 2020). La metodología corresponde a un recorrido sobre las técnicas que proponemos para lidiar con el desbalance de CLEF, comparando sus aplicaciones en variantes supervisadas con el dataset completo vs. utilizando *Active Learning*. Para la evaluación del desempeño usamos el Reporte de Clasificación de *Scikit-Learn* enfocándonos en el impacto sobre la clase relevante, complementadas con *Area Under the Curve* (AUC), *Normalized Discounted Cumulative Gain* (NDCG) y *Label Ranking Average Precision Score* (LRAPS) para medir el rendimiento general (Pedregosa et al., 2011).

2.4 Representation through masking

Masked language modeling is the task of masking **TOKENS** in a sequence and asking the method to fill that mask with an appropriate token. To complete the activity, we used the BERT model of encoding representations with transformers (Devlin et al., 2018).

Yuan proposes to enhance the selection by uncertainty in the Active Learning model through masking (Yuan et al., 2020). Based on this approach, random masking is applied to sentences with Fill-Mask, achieving a lighter representation.

2.5 Traditional Supervised Classification and Active Learning

For this experiment, we focus on two methods: Random Forest and Logistic Regression (Carvallo et al., 2020). The methodology consists of a survey of the techniques that we propose to deal with the imbalance in CLEF, comparing its applications in supervised variants with the complete dataset vs. using Active Learning. For performance evaluation we use the Scikit-Learn Classification Report, focusing on the impact on the relevant class, supplemented by Area Under the Curve (AUC), Normalized Discounted Cumulative Gain (NDCG), and Label Ranking Average Precision Score (LRAPS) to determine overall performance.

3. RESULTADOS Y DISCUSIÓN

3.1 Experimentación sobre el Dataset

La experimentación con *data augmentation* produjo 10.000 nuevos pares relevantes, alcanzado a una proporcionalidad de 94,6% y 5,3% de no relevantes y relevantes respectivamente. De igual forma se aplicó el método de enmascaramiento aleatorio. En la aplicación de los *Embeddings* se utilizó el codificador Word2Vec de Gensim con modelo CBOW. Por su parte, para BERT fue descargado el modelo *Bert-Tokenizer* de *Hugging Face*, en su versión 'uncased' con palabras en minúscula.

3.2 Clasificación Supervisada Tradicional

Usamos *Random Forest* con 300 estimadores y criterio 'gini', junto a Regresión Logística con un algoritmo de optimización 'liblinear' con 1.000 iteraciones, ambos de la librería Scikit-Learn (Pedregosa et al., 2011). Para los ensayos, el 75% de los datos se usaron para entrenamiento y 25% para evaluar. En la **Tabla 1** son resumidos los resultados más relevantes.

3. RESULTS AND DISCUSSION

3.1 Experimentation on the Dataset

The experimentation with data augmentation produced 10,000 new relevant pairs, reaching a proportionality of 94.6% and 5.3% of not relevant and relevant, respectively. The random masking method was applied. For the application of Embeddings, the Gensim Word2Vec encoder with CBOW model was used. On the other hand, for BERT the Hugging Face Bert-Tokenizer model was downloaded in its 'uncased' version with lowercase words.

3.2 Traditional Supervised Classification

We use Random Forest with 300 estimators and 'gini' criteria, together with Logistic Regression with a 'liblinear' optimization algorithm with 1,000 iterations, both from the Scikit-Learn library (Pedregosa et al., 2011). During trials, 75% of the data was used for training and 25% for evaluation. **Table 1** summarizes the most relevant results.

Model	*Embedding*	*Precision*		*Recall*		*f1-Score*		*AUC*	*NDCG*	*LRAPS*
		NoRel	*Rel*	*NoRel*	*Rel*	*NoRel*	*Rel*			
Log Reg	W2Vec	0,98	0,62	0,91	0,88	0,94	0,72	0,797	0,966	0,954
RF	W2Vec	0,99	0,78	0,95	0,95	0,97	0,85	0,883	0,981	0,974
Log Reg	W2Vec + Aug	0,83	0,62	0,97	0,18	0,89	0,28	0,575	0,931	0,907
RF	***W2Vec + Aug***	***0,95***	***0,96***	***0,99***	***0,78***	***0,97***	***0,86***	***0,886***	***0,981***	***0,974***
Log Reg	BERT	0,98	0,05	0,99	0,01	0,99	0,01	0,501	0,993	0,991
RF	BERT	0,99	0,90	0,99	0,13	0,99	0,23	0,566	0,994	0,992
Log Reg	BERT + Aug	0,95	0,47	0,99	0,01	0,97	0,01	0,503	0,981	0,973
RF	***BERT + Aug***	***0,98***	***0,99***	***0,99***	***0,72***	***0,99***	***0,83***	***0,861***	***0,994***	***0,993***
Log Reg	BERT+Aug+Mask	0,95	0,08	0,99	0,01	0,97	0,01	0,500	0,981	0,973
RF	BERT+Aug+Mask	0,95	0,99	0,99	0,07	0,98	0,14	0,537	0,982	0,976

Tabla 1. Desempeño de los modelos de clasificación supervisado sobre el *dataset* completo, aplicando combinaciones de las metodologías propuestas.

Table 1. Performance of the supervised classification models on the complete dataset, applying combinations of the proposed methodologies.

Se aprecia que *Random Forest* tiene mayor constancia, entregando buenas predicciones de los pares relevantes. Por otro lado, Regresión Logística consigue resultados cercanos, considerando un tiempo de entrenamiento menor, concordando así con otros trabajos (Carvallo et al., 2020).

Se observa que la mejor generalización es posible con BERT y DA con AUC de 0,861 siendo el que mayor f1-score posee sobre la clase relevante con 0,86. Le sigue Word2Vec y *data augmentation* con 0,886 en AUC junto a un 0,86 de f1-score en relevante. Así, ayudamos en la generalización de las clases a través del uso de nuestra técnica de *data augmentation*. A pesar de esto, cabe mencionar que BERT requiere de un mayor tiempo de entrenamiento y memoria.

3.3 Clasificación con *Active Learning*

Para *Active Learning* usamos modAL (Danka & Horvath, 2018), con incertidumbre de Shannon para consultar al ORÁCULO (Freund et al., 1997). Usamos *Random Forest* y Regresión Logística sobre las técnicas propuestas para un manejo de la base desproporcionada. Inicialmente se entrena el modelo en 2,5% de los datos. Se realizan 250 iteraciones, consultando al oráculo por la etiqueta de un dato en cada iteración para reentrenar el modelo. Los rendimientos destacados aparecen en la **Figura 3.**

It is noted that Random Forest has greater constancy, providing good predictions of the relevant pairs. Alternatively, Logistic Regression achieves close results, considering a shorter training time, thus agreeing with other studies (Carvallo et al., 2020). Our proposal achieves a better recall on the relevant pairs with respect to previous works, as a result of managing the disproportionality of the dataset.

Results show that the best generalization is possible with BERT and data augmentation, with an AUC of 0.861, reaching the highest f1-score over the relevant class with 0.86. It is followed by Word2Vec and data augmentation with 0.886 in AUC along with 0.86 in relevant f1-score. Thus, we help in the generalization of the classes through the use of our data augmentation technique. In spite of this, it should be pointed out that BERT requires a longer training time and more memory.

3.3 Classification with Active Learning

In the case of Active Learning, we used modAL (Danka et al., 2018) with Shannon's uncertainty to consult the ORACLE (Freund et al., 1997). Random Forest and Logistic Regression were used on the proposed techniques for disproportionate base management. Initially, the model was trained on 2.5% of the data. 250 iterations are performed, consulting the oracle for the label of a piece of data in each iteration to retrain the model. Highlighted performances are shown in **Figure 3.**

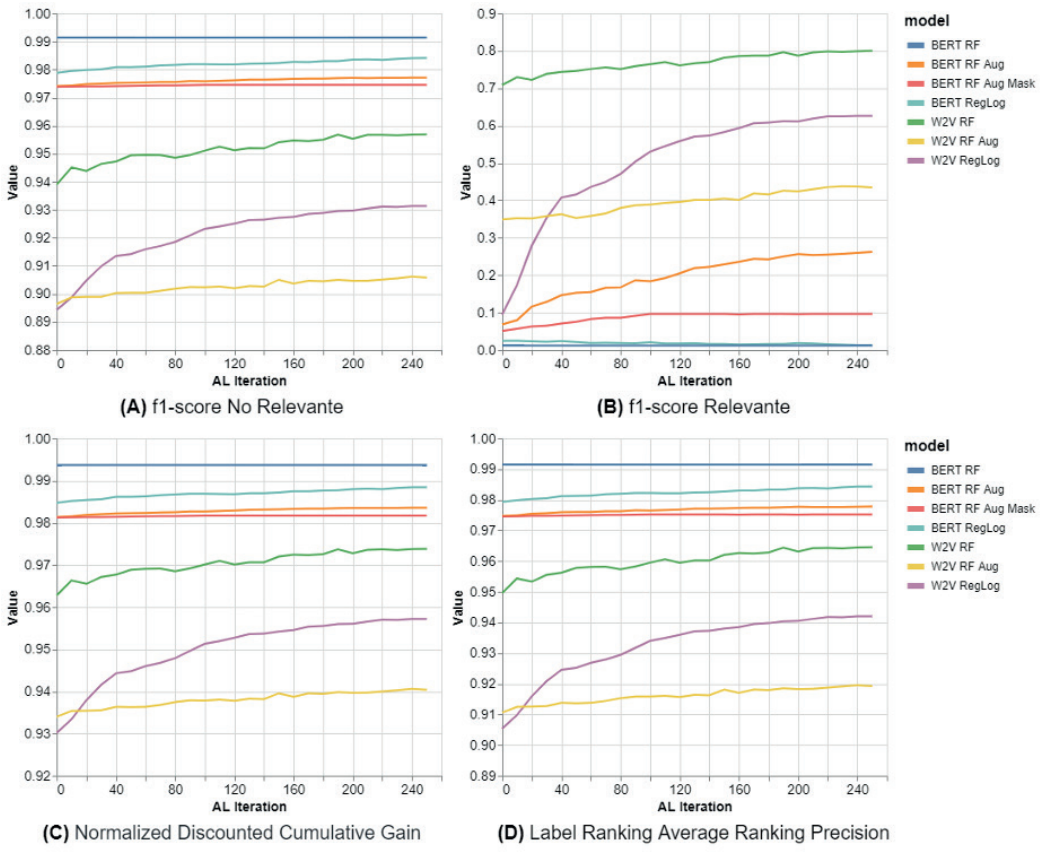

Figura 3. Progreso de las combinaciones de modelos con las técnicas propuestas durante el entrenamiento con *Active Learning* con 250 iteraciones de consultas al oráculo.

Figura 3. Progress of the combinations of models with the techniques proposed during training with Active Learning through 250 iterations of queries to the oracle.

Existe un impacto positivo con cada consulta al oráculo en modelos que utilizan las técnicas propuestas, a diferencia de BERT simple que no consigue mejora alguna sobre el f1-score de las clases relevantes.

There is a positive impact with each query to the oracle in models that use the proposed techniques, in contrast to simple BERT that does not achieve any improvement of the f1-score for the relevant classes.

Model	*Embedding*	*Iter.*	*Precision*		*Recall*		*f1-Score*		*AUC*	*NDCG*	*Prec. Rank.*
			NoRel	*Rel*	*NoRel*	*Rel*	*NoRel*	*Rel*			
Log Reg	W2Vec	0	0,82	0,96	0,99	0,10	0,90	0,18	0,551	0,934	0,911
Log Reg	W2Vec	250	0,89	0,82	0,97	0,53	0,93	0,64	0,748	0,957	0,942
RF	W2Vec	0	0,92	0,81	0,96	0,64	0,94	0,71	0,801	0,963	0,950
RF	**W2Vec**	**250**	**0,94**	**0,89**	**0,98**	**0,73**	**0,96**	**0,80**	**0,853**	**0,973**	**0,964**
Log Reg	W2Vec + Aug	0	0,80	0,56	0,99	0,01	0,89	0,01	0,501	0,926	0,901
Log Reg	W2Vec + Aug	250	0,80	0,45	0,99	0,01	0,89	0,01	0,504	0,926	0,900
RF	*W2Vec + Aug*	*0*	*0,83*	*0,68*	*0,98*	*0,20*	*0,90*	*0,31*	*0,589*	*0,934*	*0,911*
RF	***W2Vec + Aug***	***250***	***0,84***	***0,72***	***0,97***	***0,27***	***0,90***	***0,39***	***0,620***	***0,938***	***0,916***
Log Reg	BERT	0	0,98	0,02	0,97	0,03	0,98	0,02	0,502	0,984	0,979
Log Reg	BERT	250	0,98	0,01	0,99	0,01	0,98	0,01	0,498	0,988	0,984
RF	BERT	0	0,98	0,86	0,99	0,01	0,99	0,01	0,502	0,993	0,991
RF	BERT	250	0,98	0,99	0,99	0,01	0,99	0,01	0,503	0,994	0,991
Log Reg	BERT + Aug	0	0,95	0,12	0,97	0,07	0,96	0,09	0,522	0,972	0,962
Log Reg	BERT + Aug	250	0,95	0,13	0,98	0,06	0,96	0,08	0,518	0,975	0,965
RF	*BERT + Aug*	*0*	*0,95*	*0,95*	*0,99*	*0,04*	*0,97*	*0,07*	*0,518*	*0,981*	*0,974*
RF	***BERT + Aug***	***250***	***0,96***	***0,98***	***0,99***	***0,18***	***0,98***	***0,31***	***0,591***	***0,984***	***0,978***
Log Reg	BERT+Aug+Mask	0	0,95	0,06	0,98	0,02	0,97	0,03	0,501	0,975	0,966
Log Reg	BERT+Aug+Mask	250	0,95	0,05	0,99	0,01	0,97	0,01	0,499	0,978	0,971
RF	BERT+Aug+Mask	0	0,95	0,99	0,99	0,04	0,98	0,07	0,517	0,981	0,974
RF	BERT+Aug+Mask	250	0,95	0,99	0,99	0,05	0,97	0,10	0,525	0,981	0,975

Tabla 2. Desempeño de los modelos de clasificación entrenados con *Active Learning*.

Table 2. Performance of the classification models including training using Active Learning.

De acuerdo con la **Tabla 2**, el uso de AL alcanza buenos resultados de generalización. Para W2V con RF, acabamos obteniendo un f1-score de 0,8 sobre relevantes y AUC de 0,853. Este progreso es bastante destacable, siendo cercano al obtenido en la sección tradicional. Por su parte, para BERT, nuestra metodología aumenta su f1-score de 0,07 a 0,31 sobre relevantes, obteniendo 0,591 de AUC.

En *Active Learning* no logramos alcanzar el mismo nivel de resultados usando BERT que en trabajos relacionados (Carvallo et al., 2020), mas sí conseguimos una señal positiva en Word2Vec con el uso de DA, logrando 0,39 de f1-score en relevantes con solo 2,6% del dataset.

As shown in **Table 2**, the use of Active Learning achieves good generalization results. For W2V with RF, we ended up obtaining an f1-score of 0.8 on relevant classes and an AUC of 0.853. This progress is quite remarkable, being close to that obtained in the traditional section. As for BERT, our methodology increases its f1-score from 0.07 to 0.31 on relevant classes, reaching 0.591 of AUC.

In Active Learning we were not able to achieve the same level of results using BERT as in related works (Carvallo et al., 2020), although we did achieve a positive signal in Word2Vec with the use of data augmentation, achieving 0.39 of f1-score in relevant classes with only 2.6% of the dataset.

4. CONCLUSIONES

Por medio de la investigación realizada fue posible experimentar con una serie de técnicas para clasificar involucrando Embeddings de documentos, junto con proponer métodos de data augmentation y enmascaramiento para combatir la desproporcionalidad del dataset CLEF.

Además, fue posible explorar *Active Learning*, que mejoró los resultados a partir de una cantidad menor de datos etiquetados. Esto corrobora nuestra hipótesis de que el modelo es capaz de seleccionar los mejores datos para entrenar. Lo anterior se reflejó con Word2Vec en *Random Forest*, alcanzando 0,853 de AUC y 0,8 en f1-score en relevantes con solo 2,6% del *dataset*. De igual forma, el mejor impacto de nuestro *Data Augmentation* fue con RF al obtener 0,39 de f1-score en relevantes. Finalmente, el mayor efecto provocado por nuestros métodos fue para *Random Forest* con BERT pasando de 0,08 a 0,31 en f1-score sobre relevantes, alcanzando 0,591 de AUC. Por su parte, en clasificación tradicional, BERT y *data augmentation* llegó a 0,861 de AUC con 0,83 de f1-score en relevantes. Le siguió Word2Vec y DA con AUC de 0,886 y un f1-score de 0,86 sobre relevantes.

Podemos afirmar que las metodologías propuestas permiten adaptar el trabajo de clasificación de literatura médica bajo condiciones de desbalance en la proporcionalidad de clases como en CLEF y a la necesidad de disminución de recursos por los costos asociados al etiquetado de los datos. Aquí el uso de *Data Augmentation* en base a *Fill-Mask* tiene un impacto positivo sobre el entrenamiento supervisado tradicional, y sobre *Active Learning* con una cantidad limitada de datos etiquetados.

Como trabajo futuro se plantea el estudio de Epistemonikos que presenta recursos en diferentes idiomas y con otro nivel de tecnicismos (Rada et al., 2013). Finalmente, se sugiere la experimentación con nuevas metodologías de muestreo en *Active Learning*, incorporando medidas que incluyan información complementaria a la incertidumbre de los datos, como el uso de interés didáctico y/o innovador de los artículos relevantes que son candidatos para la búsqueda de evidencia (Lee et al., 2020).

Agradecimientos

El trabajo realizado no hubiese sido posible sin el apoyo de mi familia y amigos quienes me alentaron a probar nuevos desafíos. Agradecer en especial al profesor Denis Parra por la oportunidad que me otorgó y los consejos que me ayudaron en mi formación; de igual manera a Andrés Carvallo por orientarme durante el proceso y al grupo HAIVis Lab (antiguamente SocVis) por el grato ambiente y conocimiento nuevo que generaron durante la investigación.

4. CONCLUSIONS

Throughout this study, we experimented with a series of classification techniques involving the method of document representation, along with our proposal of data augmentation and masking methods to address the disproportionality of the CLEF dataset.

In addition, we could explore Active Learning, which improved results obtained from a smaller set of labeled data. This corroborates our hypothesis that the model is capable of selecting the best data for training, which was reflected with Word2Vec in Random Forest, reaching 0.853 in AUC and 0.8 in f1-score in relevant with only 2.6% of the dataset. In the same way, the best impact of our data augmentation was with RF, obtaining 0.39 of f1-score in relevant. Finally, the greatest effect caused by our methods was for Random Forest with BERT going from 0.08 to 0.31 in f1-score on relevant, reaching 0.591 of AUC. On the other hand, in traditional classification, BERT and data augmentation reached 0.861 of AUC with 0.83 of f1-score in relevant, followed by Word2Vec and DA with AUC of 0.886 and an f1-score of 0.86 on relevant.

We can assert that the proposed methodologies allow adapting the work of classification of medical literature under conditions of imbalance in the proportionality of classes as in CLEF and to the need of reducing resources due to the costs of labeling the data. Here, the use of data augmentation based on Fill-Mask has a positive impact on traditional supervised training, and on Active Learning with a limited amount of labeled data.

As future work, we propose the analysis of Epistemonikos, which has resources in different languages and with another level of technicalities. Finally, we suggest experimentation with new sampling methodologies in Active Learning, incorporating measures that include information complementary to the uncertainty of the data, such as the didactic and/or innovative use of relevant articles that are candidates in the search for evidence.

Acknowledgements

This work would not have been possible without the support of my family and friends, who encouraged me to try new challenges. Special thanks to Professor Denis Parra for the opportunity and advice throughout my training. Many thanks to Andrés Carvallo for guiding me during the process, and the HAIVis Lab (formerly SocVis) for the pleasant atmosphere and new knowledge generated during this investigation.

GLOSARIO

EMBEDDING: En el contexto del aprendizaje automático, un Embedding es un espacio de dimensiones relativamente bajas en el que se pueden traducir vectores de dimensiones altas. Los Embeddings facilitan el aprendizaje en entradas grandes, como vectores dispersos que representan textos.
ORÁCULO: En el contexto de aprendizaje automático, se refiere a usuarios humanos del sistema que pueden responder o etiquetar cierto elemento en particular en un conjunto de datos. En este trabajo se asume que el oráculo responderá siempre de forma correcta.
TOKEN: En el contexto de aprendizaje automático, se refiere a uno de los elementos que representan un texto, correspondiendo generalmente a una palabra de una oración. Un derivado es la Tokenización, que consiste en la separación de una oración en tokens o palabras tratadas como elementos individuales.

GLOSSARY

EMBEDDING: In the context of machine learning, an Embedding is a relatively low-dimensional space into which high-dimensional vectors can be translated. Embedding facilitate learning on large inputs, such as sparse vectors representing texts.
ORACLE: In the context of machine learning, it refers to human users of the system who can reply to or label a certain item in a data set. In this work it is assumed that the oracle will always answer correctly.
TOKEN: In the context of machine learning, it refers to one of the elements that represents a text, generally corresponding to a word in a sentence. Tokenization is a derivation consisting in the separation of a sentence into tokens or words treated as individual elements.

REFERENCES

Baldridge, J., & Osborne, M. (2004). Active learning and the total cost of annotation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (pp. 9-16).

Canese, K., & Weis, S. (2013). PubMed: the bibliographic database. In The NCBI Handbook [Internet]. 2nd edition. National Center for Biotechnology Information (US).

Carvallo, A., Parra, D., Lobel, H., & Soto, A. (2020). Automatic document screening of medical literature using word and text embeddings in an active learning setting. Scientometrics, 1-38.

Danka, T., & Horvath, P. (2018). modAL: A modular active learning framework for Python. arXiv preprint arXiv:1805.00979.

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association, 24(3), 596-606.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Doms, A., & Schroeder, M. (2005). GoPubMed: exploring PubMed with the gene ontology. Nucleic acids research, 33(suppl_2), W783-W786.

Ertekin, S., Huang, J., & Giles, C. L. (2007). Active learning for class imbalance problem. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 823-824).

Freund, Y., Seung, H. S., Shamir, E. y Tishby, N. (1997). Muestreo selectivo mediante el algoritmo de consulta por comité. Aprendizaje automático, 28 (2-3), 133-168.

Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., ... & Zuccon, G. (2017, September). CLEF 2017 eHealth evaluation lab overview. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 291-303). Springer, Cham.

Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. Nature, 555(7695), 175-182.

Maynard, A. (1997). Evidence-based medicine: an incomplete method for informing treatment choices. The Lancet, 349(9045), 126-128.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, July). Semantically equivalent adversarial rules for debugging nlp models. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 856-865).

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't.
Settles, B. (2009). Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences.
Settles, B. (2011). From theories to queries: Active learning in practice. In Active Learning and Experimental Design workshop In conjunction with AISTATS 2010 (pp. 1-18).
Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243.
Yuan, M., Lin, H. T., & Boyd-Graber, J. (2020). Cold-start active learning through self-supervised language modeling. arXiv preprint arXiv:2010.09535

EQUIPO DE INVESTIGADORES / RESEARCH TEAM

Benjamín Ayancán

Denis Parra