

Clasificación de artículos científicos relacionados a la agricultura de precisión con imágenes de UAV por medio de técnicas de aprendizaje profundo

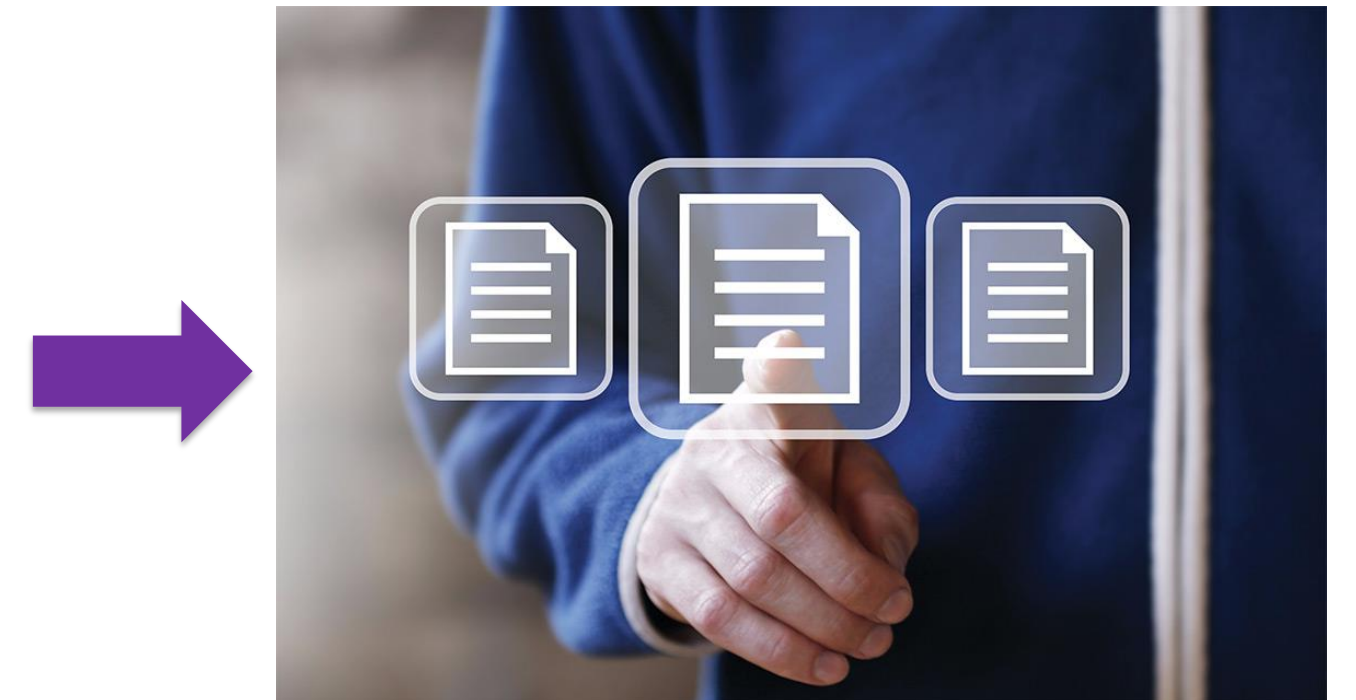
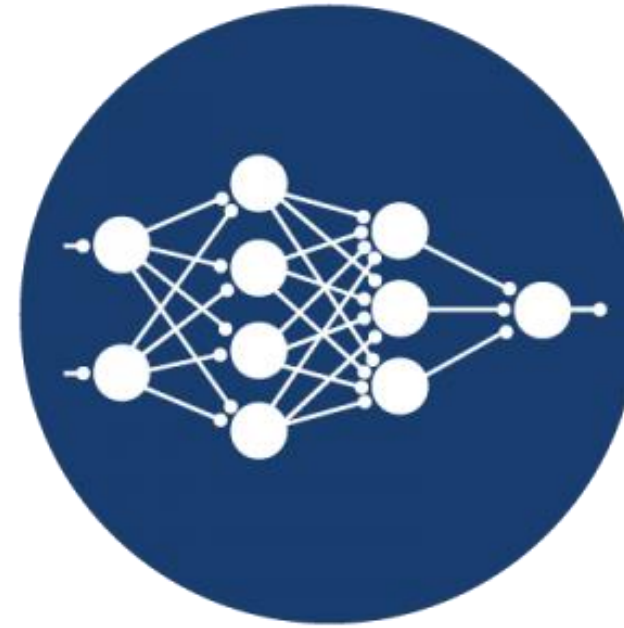
PROBLEMA

La revisión literaria es el punto de partida para el desarrollo de un trabajo académico. Esta tarea se realiza de forma manual, siendo tediosa y extensiva; puesto que sólo se puede dictaminar el nivel de relevancia de los artículos científicos enmarcados en el tópico deseado, a través de la lectura de los mismos.



OBJETIVO GENERAL

Crear un modelo que clasifique artículos científicos en inglés relacionados a la agricultura de precisión con imágenes UAV.

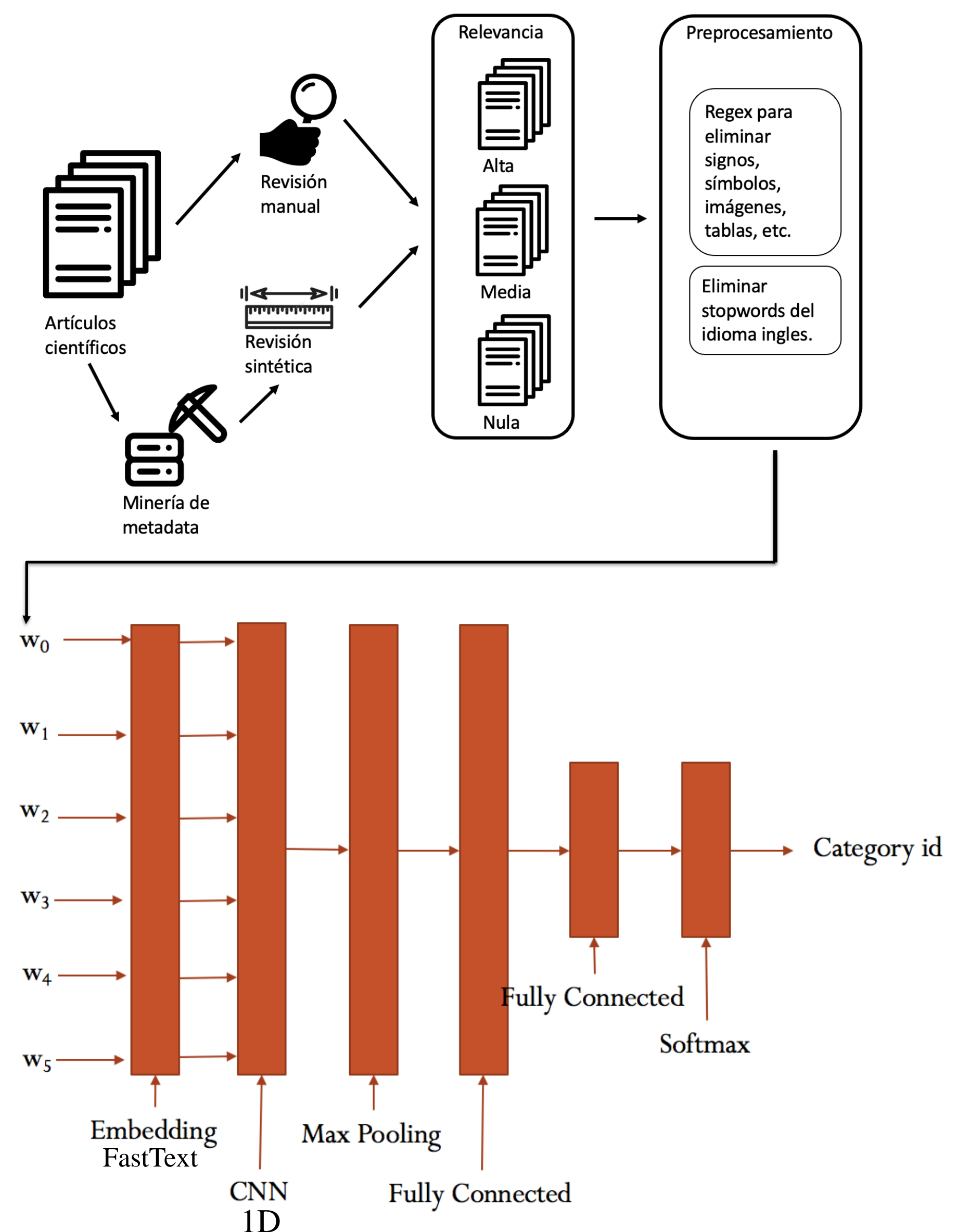


PROPUESTA

Generar un modelo aprendizaje profundo que aprenda de manera supervisada, a clasificar artículos científicos relacionados a la agricultura de precisión con imágenes de UAV, en 3 categorías: alta, media y nula relevancia. A fin de optimizar el proceso de revisión y selección de las fuentes bibliográficas a utilizar en una investigación académica.

La propuesta se divide en 4 módulos:

1. Adquisición del conjunto de datos: se tomó artículos científicos de repositorios provistos por CIB-ESPOL y se los etiquetó manual y sintéticamente en las 3 clases mencionadas.
2. Minería de metadata: para la conformación de una métrica Ad hoc que clasifique sintéticamente a los documentos, se minó información de los mismos, a través de la API de Crossref.
3. Representación del texto: los corpus de los documentos pasaron por un flujo de preprocesamiento estándar y luego de vectorización con el modelo FastText, para componer la capa de embedding del modelo de aprendizaje profundo.
4. Clasificación del texto: la capa obtenida con la vectorización es la entrada para un modelo CNN 1D, de optimizador Adam, LR 0.001 y una función de pérdida de BCE; para clasificar los artículos acorde a los niveles de relevancia planteados.



RESULTADOS

1. El mejor modelo, se dio a partir del entrenamiento sobre el conjunto de datos de 84 documentos, etiquetados de la siguiente manera: 30 para alta, 27 de media y 27 con nula relevancia; a través, del etiquetado manual de 40 artículos y por contraparte la clasificación de 44 piezas académicas mediante un etiquetado sintético que utilizó la métrica Ad hoc.
2. El modelo clasificador llegó a tener un F1 score del 87 % en pruebas con 18 ejemplos inéditos (6 por cada clase). Lo que se refleja en la matriz de confusión entre sus niveles, donde la diagonal presenta valores cercanos a 33%, que sería el valor ideal.
3. Se desarrolló un prototipo de aplicación web, con la finalidad de crear una herramienta en la que se pueda realizar la clasificación de los trabajos académicos, por medio del modelo generado, en pro de poder hacer que sea accesible para cualquier investigador y así lograr visualizar los resultados del proceso de categorización.

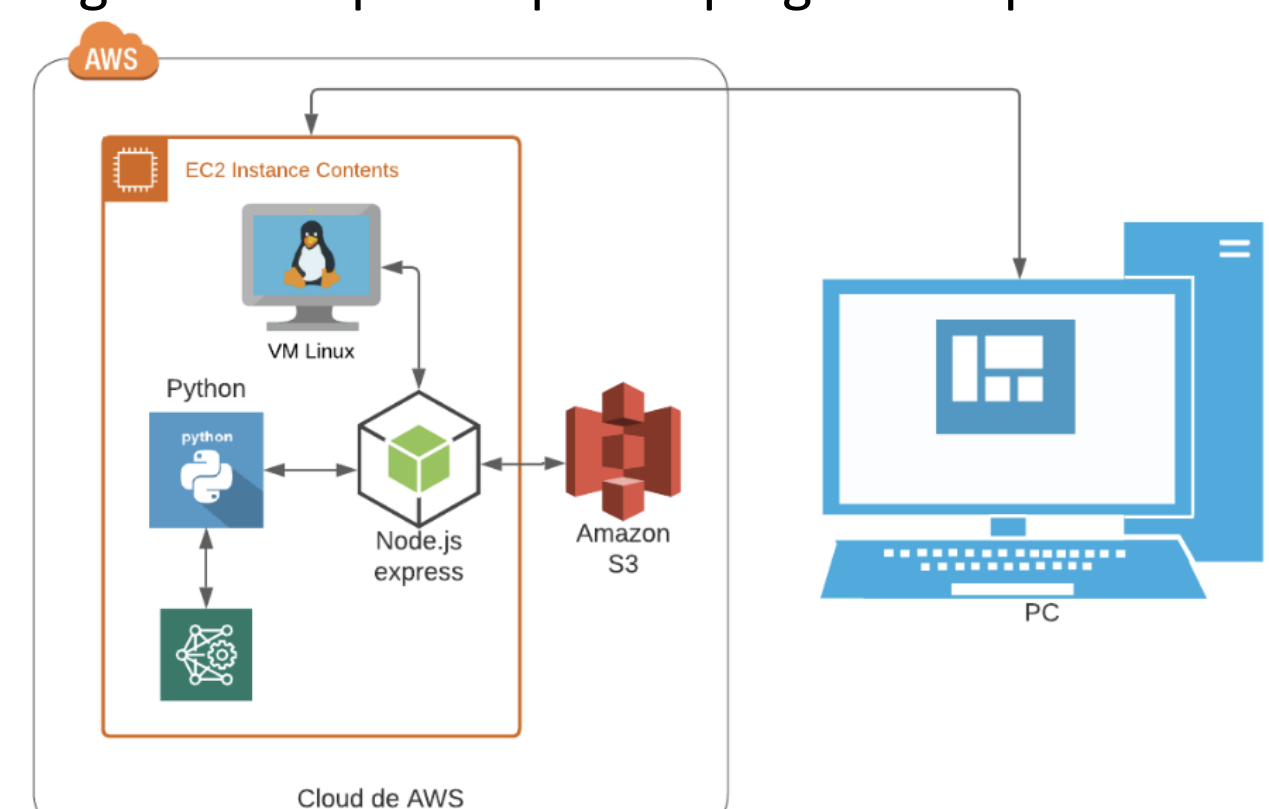
Métricas del mejor modelo alcanzado

	Precisión	Exhaustividad	F1 score
Alta	0,88	0,86	0,87
Media	0,80	0,89	0,84
Nula	0,86	0,94	0,90
Media	0,84	0,89	0,87

Matriz de confusión del mejor modelo alcanzado

		Predicción		
		Alta	Media	Nula
Real	Alta	27,38 %	5,95 %	2,38 %
	Media	2,38 %	27,39 %	2,38 %
	Nula	1,19 %	2,38 %	28,57 %

Diagrama del prototipo desplegado en producción



CONCLUSIONES

- Se logró entrenar un modelo que los discriminara en 3 niveles acorde a su relevancia, que son alto, medio y nulo. En el cual se obtuvo métricas de 94% en entrenamiento para F1 score como promedio para todas sus categorías y para prueba obtuvo un 87% en la métrica indicada.
- A partir del conjunto de datos etiquetados, se diseñó un flujo de procesos que permitan la transformación de los corpus de los documentos, para su óptima vectorización y posterior clasificación.
- Se creó una métrica Ad hoc para completar el etiquetado manual, de forma que sintéticamente se aumentó la cantidad de artículos científicos etiquetados, mejorando la calidad en los entrenamientos del modelo.
- Se implementó un prototipo el cual puede ser ejecutado desde cualquier parte y corroborar los resultados obtenidos por cuenta propia.