Asistencia para la categorización de documentos

Álvaro Moncada – alvaro.moncada @javeriana.edu.co
Álvaro Pérez – perez.alvaro @javeriana.edu.co
Edwin Turizo – edwin.turizo@javeriana.edu.co

Resumen. El presente documento tiene como intención mostrar una problemática en el campo del E-Discovery en la asistencia para la categorización de documentos de forma que se le delegue esta actividad a un sistema inteligente basado en agentes racionales que puedan utilizar múltiples técnicas de inteligencia artificial para la clasificación y/o segmentación de textos. Entre estas técnicas se plantean las redes neuronales, lógica difusa, métodos evolutivos y aprendizaje de máquina que dependiendo al tipo de aprendizaje que se requiera, se tendrán distintos procesos de tratamiento de los datos de entrada y los posibles a obtener.

Palabras Clave: Clasificación, Aprendizaje de Máquina, Redes Neuronales, Lógica Difusa, Algoritmos Genéticos, Sistemas Inteligentes.

1. Motivación y Contextualización del Problema

El descubrimiento de evidencia digital o E-Discovery, permite a investigadores o abogados identificar información almacenada electrónicamente relacionada con el objetivo de un caso. Luego de identificar aquellos archivos relevantes inicia la etapa de revisión que, usualmente, consume muchos recursos y toma mucho tiempo debido al gran volumen de información a revisar, por ejemplo, para que una persona revise 10.000 documentos podría estar tardando 41 días, suponiendo 2 minutos por documento y una jornada de 8 horas laborales. Por lo que es de importancia la implementación de técnicas que reduzcan estos tiempos y permitan a los revisores enfocar sus esfuerzos para identificar la mayor cantidad de información relevante ágilmente.

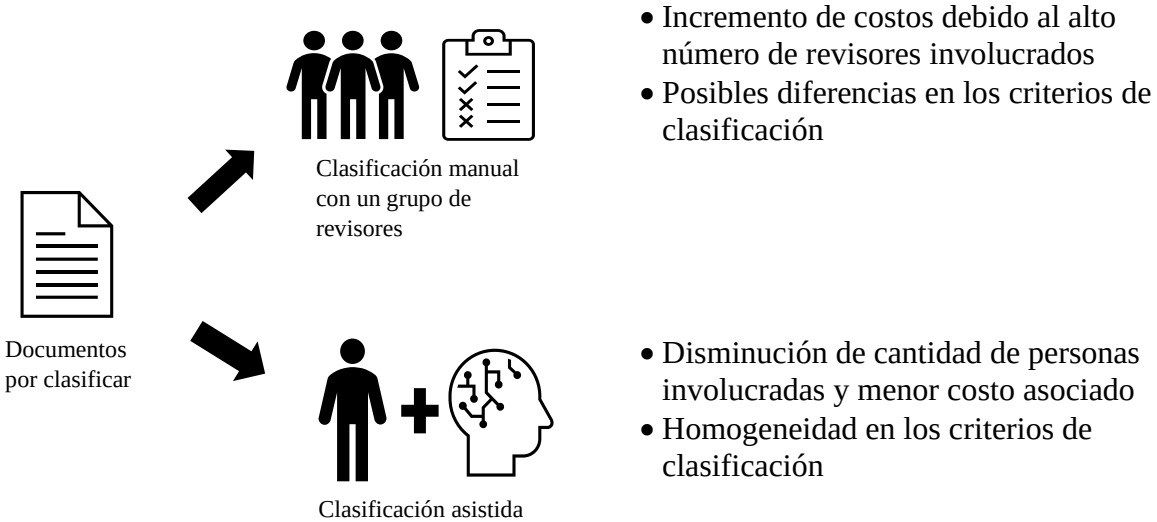

Ilustración 1. Comparación de tipos de revisión de documentos

En la actualidad existen múltiples herramientas que facilitan la revisión de esta información (e.g. Relativity, Brainspace, Reveal), las cuales ya implementaron técnicas de IA para asistir en la revisión de documentos, las cuales, entre otros términos, han acuñado el nombre de revisión asistida por tecnología (*Technology Assisted Review*).

Estas técnicas comprenden, pero no se limitan, a [1]:

- **Unsupervised learning algorithms**: Algoritmos que infieren categorías de documentos similares sin entrenamiento por parte de un experto en la materia. Por ejemplo, Clustering, Near-Duplicate Detection, Concept search.
- **Supervised learning algorithms**: Algoritmos para organizar o clasificar documentos mediante análisis de sus características basándose en la codificación brindada en un training set o muestra por parte del revisor. Por ejemplo, Support vector machine, Logistic regression, Bayesian classifiers.

A partir de estas técnicas, la problemática a abordar será la clasificación automática de documentos basándonos en técnicas de Aprendizaje de Máquina, Redes Neuronales, Lógica Difusa, Algoritmos Genéticos

2. Descripción de la Tarea

2.1. Visión General

La tarea consiste en diseñar e implementar un sistema multiagente que pueda realizar la correcta categorización de documentos de texto, estableciendo la temática de la que trata el texto por medio de diversas técnicas de IA.

2.2. Caracterización Agente

El proyecto debe contar con agentes racionales de forma que se tenga una entidad con la meta de categorizar el tema de un texto. Estos agentes se desenvuelven en un ambiente que percibe las palabras y en base a su frecuencia de repetición, se genera un modelo de decisión influenciado por las validaciones humanas de los resultados y que actúa de forma autónoma para la clasificación y/o segmentación del texto. La proactividad del agente se evidencia en el decidir y actuar en el proceso sin la intervención directa de la persona por la distinción de la temática y poder realizar el procesamiento y análisis de las variables de entrada.

2.3. Restricciones y Alcances

- Se tendrán documentos de texto de noticias para la categorización de la temática.
- Los datos que se necesiten para el desarrollo del proyecto serán de uso abierto y de dominio público.
- El alcance está limitado por los tiempos dispuestos para cada una de las fases y entregas del proyecto.
- Al no tenerse un cliente directo, la validación de los modelos, resultados y pruebas serán realizadas de forma experimental.

- Debido a las amplias temáticas de las categorías que se tienen en las noticias, se requiere reducir el número de categorías con el fin de que contribuyan al estado del arte del problema.

3. Análisis de Potencialidades para el Uso de Herramientas de Inteligencia Artificial

A continuación, se presentan antecedentes de aplicación de las técnicas IA que se han aplicado en los diversos módulos de sistemas inteligentes.

3.1. Redes neuronales

Las redes neuronales han sido utilizadas para la asistencia en la clasificación de documentos utilizando métodos supervisados, siendo ésta la técnica que mejor resultados ha dado. Para el Procesamiento de Lenguaje Natural ("NLP" por sus siglas en inglés) las redes neuronales se han utilizado con éxito en tareas como la extracción de información relevante de textos, sin tener que especificar características de un dominio en particular [4]. Algunos algoritmos utilizados:

- Convolutional Neural Networks – CNN
- Recurrent Neural Networks – RNN
- Supreme Court Classifier – SCC (mejores resultados)

3.2. Métodos evolutivos

En cuanto a los métodos evolutivos, en [5] se utilizó un algoritmo genético es utilizado junto con LSI (*Latent Semantic Features*) para optimizar el vector de los descriptores seleccionados. Para este caso se utilizaron tres *datasets* "Reuters", "Ohsumed" y "Enron", en los tres casos la precisión del algoritmo y superar a aquellos casos en los que solo se utilizó LSI. A partir de este documento podemos obtener datos sobre el marco general para dar solución a nuestra problemática, a pesar de que se utilizó con *datasets* diferentes, el propuesto consiste en tres etapas principalmente:

- Feature Extraction
- Feature selection
- Classification

3.3. Lógica difusa

Se utiliza como clasificador de texto a través de la prueba en conjuntos de documentos que busca mejorar a través de la retroalimentación continua para alcanzar la inteligencia y autonomía cognitiva hasta lograr los objetivos previstos [2]. Se habla de lógica difusa puesto que hay una alta incertidumbre y ambigüedades en la descripción y comprensión del lenguaje natural. La lógica difusa apoyara en el reto de clasificar los documentos a través del algoritmo FKOCNN y TGSOM que son técnicas basadas en clustering difuso los cuales a través de árboles de decisión ayudan a identificar las palabras que caracterizan o permiten diferenciar entre un grupo y otro. Es posible apoyarse con otros algoritmos como Latent semantic analysis (LSA) y Vector space model(VSM) para mejorar el proceso de clustering.

3.4. Aprendizaje de Máquina

Por el lado del aprendizaje de máquina se encontraron diversos artículos, los cuales trataban de encontrar un enfoque de machine learning por medio de diversos algoritmos, con el fin de clasificar

documentos electrónicos tanto científicos como educacionales y buscaban establecer la temática teniendo como base atributos propios del texto [1].

Antes de realizar la categorización automática de textos se debe realizar la extracción de características, la cual consiste en obtener los atributos que mejor describan el texto a clasificar, esto es, transformar el archivo de texto de forma que se pueda tener como entrada al método de aprendizaje de máquina [2].

A continuación, se presentan los algoritmos y/o métodos que se han aplicado y que actualmente se tienen en cuenta como opciones de implementación.

Support Vector Machine (SVM)

Las máquinas de vectores de soporte (SVM) son modelos de aprendizaje supervisado que analizan los datos utilizados para el análisis de clasificación y regresión. Una SVM es un clasificador que intenta maximizar el margen entre los datos de entrenamiento y el límite de clasificación mediante la construcción de un hiperplano o conjunto de hiperplanos. La idea es maximizar el margen aumentando la posibilidad de que la clasificación sea correcta en nuevos datos [3].

K-Means

El algoritmo K-Means es un algoritmo de aprendizaje no supervisado que busca agrupar elementos en grupos basándose en sus características similares. La forma de aplicación en textos varia puesto que, primeramente, se debe crear una matriz de términos frecuentes por medio de la medida TF-IDF, que expresa la relevancia de cada termino en un documento por medio del número de ocurrencias. Posteriormente, esa matriz sirve de entrada al algoritmo K-Means para agrupar los documentos en base a sus similitudes estableciendo los vecinos cercanos y así tener categorías definidas de los documentos relacionados [1].

4. Análisis de Viabilidad de Validación Experimental

Utilizaremos una base de datos gratuita, la cual contiene alrededor de 25 mil noticias desde 2012 hasta 2018 obtenidas de Huffpost [link], en dicha base se encuentra el texto de diferentes noticias y una clasificación con respecto al tema al que hace referencia el texto. Las variables controlables en este caso son la descripción de las noticias y el encabezado las cuales se realizarán métodos de pre y procesamiento como filtrar la base por noticias de política, negocios, impacto, divorcio, crimen, noticias mundiales, tecnología dinero y educación. Además del uso de *tokenizacion* dividir las cadenas de texto de la variable noticias o encabezado de noticias para posteriormente usar *bag of words* de Python que busca extraer características relevantes o que puedan influenciar en definir a qué tipo de noticia pertenece la variable para de esta manera apoyar la clasificación.

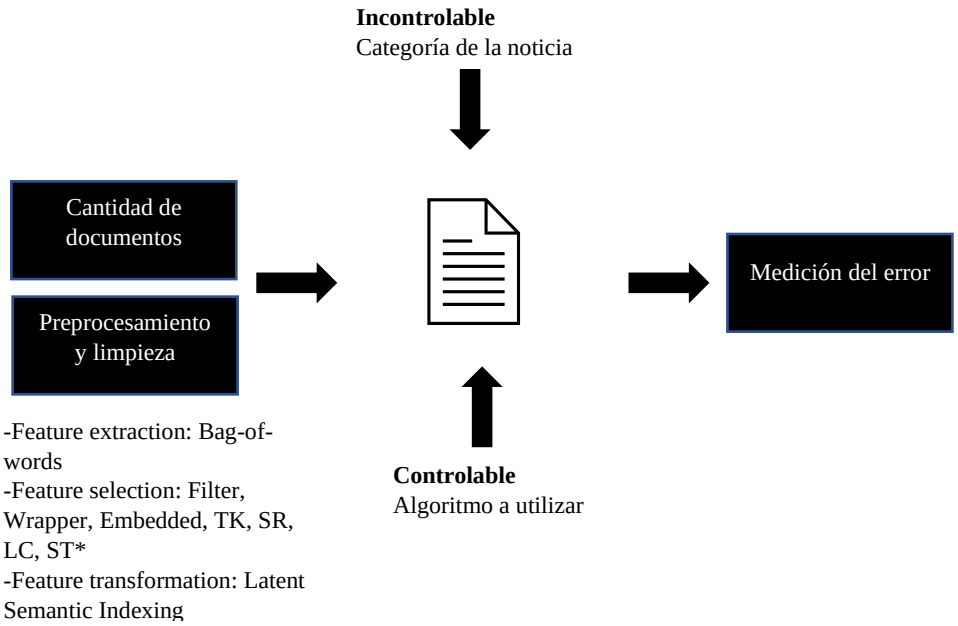

Ilustración 2. Esquema experimental para la clasificación automática de documentos
*TK:Tokenization, SR: Stop-word removal, LC: Lower case conversion, ST: Stemming

Referencias bibliográficas

[1] W. Yu and X. Linying, "Research on Text Categorization of KNN Based on K-Means for Class Imbalanced Problem," 2016 Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), Harbin, China, 2016, pp. 579-583, doi: 10.1109/IMCCC.2016.61.

[2] NADIA ARAUJO ARREDONDO, Método semisupervisado para la clasificación automática de textos de opinión, 2009, Disponible en: inaoe.repositorioinstitucional.mx/jspui/handle/1009/365

[3] Martha Vásquez Moo, Víctor Cetina, Carlos Brito, “Clasificación de documentos usando Máquinas de Vectores de Apoyo”, 2012

[4] Samir Undavia, Adam Meyers, John E. Ortega “A Comparative Study of Classifying Legal Documents with Neural Networks.” IEEE Xplore 2018

[5] Alper Kursat Uysal, Serkan Gunal, “Text classification using genetic algorithm oriented latent semantic features” Science Direct, 2014.