Clasificación y depuración de datos de la segunda sección del Boletín Oficial de la República Argentina mediante aprendizaje de máquina

Nestor A. Balich, Franco A. Balich, Hugo Fraga

CAETI - Centro de Altos Estudios en Tecnología Informática
Universidad Abierta Interamericana. Informática (UAI)
Montes de Oca 745. Ciudad Autónoma de Buenos Aires, Argentina.

{nestor.balich, francoadrian.balich}@uai.edu.ar
hfraga@boletinoficial.gob.ar

Resumen. La segunda sección del Boletín Oficial de la República Argentina en donde se publican los avisos comerciales y judiciales es un importante medio de difusión de información para empresas, instituciones y particulares, pero la gran cantidad de información almacenada desde su creación suma a la que se publica diariamente hace que la depuración de las bases de datos sea un proceso complejo y costoso. La aplicación de técnicas de machine learning (ML) en la clasificación de textos ha evolucionado significativamente, especialmente en el uso de modelos de aprendizaje profundo en áreas como la identificación de noticias falsas y la detección de spam. También se han realizado estudios sobre el uso de técnicas de ML en la identificación y corrección de errores en bases de datos, incluyendo la corrección de errores en imágenes médicas y la identificación de avisos comerciales ilegales en la web. Se propone el desarrollo de un modelo de aprendizaje de máquina para la clasificación, detección y corrección de los avisos comerciales de la segunda sección del Boletín Oficial de la República Argentina. La comparación y evaluación de diferentes modelos de IA. La creación de dos prototipos mediante metodologías agiles de desarrollo en base al diseño de dos productos mínimos viables (MPV) que permitan rápidamente a los usuarios finales testear la usabilidad, efectividad de los prototipos y definir tiempos estimados del proceso de corrección de todos los avisos de la 2da clasificados para evaluación por parte del modelo de IA.

Palabras clave: clasificación, inteligencia artificial, aprendizaje supervisado, depuración, base de datos.

1 Introducción

Esta línea de I+D forma parte de los proyectos radicados en el Centro de Altos Estudios en Tecnología informática (CAETI) de la Universidad Abierta Interamericana (UAI). En este proyecto participan docentes, alumnos e investigadores enmarcado en

los proyectos de transferencia tecnología del laboratorio de robótica física e inteligencia artificial (LRFIA). Dentro de las líneas de investigación sobre inteligencia de los proyectos con financiamiento y duración a 2 años.

La segunda sección del Boletín Oficial de la República Argentina es un importante medio de difusión de información para empresas, instituciones y particulares. Sin embargo, la cantidad de información que se publica diariamente en esta sección hace que la depuración de bases de datos sea un proceso complejo y costoso en términos de tiempo y recursos. En este contexto, surge la hipótesis de que es posible desarrollar un modelo de inteligencia artificial basado en ML capaz de aprender a clasificar los avisos comerciales y luego catalogarlos para obtener los avisos de manera eficiente y de forma totalmente autónoma.

La aplicación de estas técnicas en la clasificación de textos es un área de investigación en constante evolución, y su aplicación en la depuración de bases de datos no es una excepción. En particular, la utilización de modelos de aprendizaje profundo ha permitido mejorar significativamente la capacidad de clasificación de textos en diversas áreas, como la identificación de noticias falsas o la detección de spam en correos electrónicos.

En el ámbito de la depuración de bases de datos, también se han realizado estudios sobre la utilización de técnicas de ML para la identificación y corrección de errores “el uso de ML para mejorar la eficiencia y la precisión de la limpieza de datos y la consideración de los efectos de la limpieza de datos en análisis estadístico” [1].

Los problemas de calidad de los datos son complejos de resolver y existente gran cantidad de técnicas y herramientas para lograrlo [2] y que debimos evaluar al seleccionar los modelos a utilizar en el presente trabajo.

Otro estudio interesante donde se desarrolló un modelo para la detección de publicidad ilegal en vallas publicitarias basado en el aprendizaje automático (FIBAD) [3]. Permitió demostrar que los resultados obtenidos por el modelo eran capaces de identificar avisos ilegales con una alta precisión y eficiencia.

En este trabajo, se propone el desarrollo y validación de un modelo de inteligencia artificial basado en ML con aprendizaje supervisado para la clasificación y catalogación de los avisos comerciales de la segunda sección del Boletín Oficial de la República Argentina. Se espera que este modelo permita obtener los avisos de manera eficiente y autónoma, lo cual representaría un avance significativo en la depuración de bases de datos en este ámbito. *“Los datos son uno de los desafíos perennes en el análisis de datos, y no hacerlo puede resultar en análisis inexactos y decisiones poco confiables”* [4].

Se realizaron los siguientes pasos:

1) Recolección de datos.
2) Proceso de análisis exploratorio de datos (EDA), donde los datos se examinan con estadísticas descriptivas simples para determinar si existen problemas de datos, como podemos observar detalladamente en el libro “Data Science in the Database: Using SQL for Data Preparation” [5] como ser duplicación de datos,

datos nulos, problema del modelo de integridad de datos con tablas relacionales, entre los principales problemas. Continuando con la identificación de los datos representativos obtenidos de cinco tablas (id del tipo de aviso, texto del aviso, texto del tema, id de registro, fecha de ingreso, fecha de publicación, número de boletín) generando un archivo "avisos2da.csv" con un total de 770.000 registros.

3) Normalización de los datos eliminando los registros con republicaciones, pues el mismo aviso debe ser republicado por la "***ley de sociedades comerciales nº 19.550***" [6] de 1 a 10 días dependiendo del tipo de aviso. Encontramos una particularidad en el modelo de datos, pues si bien el texto del aviso es el mismo no se encuentra asociado a una tabla de republicación, sino que repite el mismo registro cambiándoles la fecha, el id del registro y un flag de aviso republicado. Para nuestro modelo de entrenamiento solo necesitamos extraer los campos de textos una sola vez, con lo cual no es correcto entrenar con hasta 10 veces el mismo aviso mientras otros avisos solo tienen una ocurrencia, pues esto perjudicaría al modelo generando un sesgo dando mayor peso al de mayor ocurrencia. Con esta salvedad el dataset se redujo a 550.000 registros.
4) Agrupación por categorías de tipo de aviso (Tabla 1), detectando algunos tipos de avisos con menos de 20 ocurrencias que fueron descartados por no ser suficientes para el entrenamiento. Fijando la cantidad de avisos representativos mínima en 450 para poder extraer una muestra aleatoria y balanceada de cada categoría. Pues *"Los conjuntos de datos desequilibrados afectan negativamente el rendimiento de los algoritmos de aprendizaje automático"* [7].

Tabla 1. Resultado de agrupación por categorías.

Rubro nombre	Cantidad
Avisos comerciales	180641
Balances	527
Citaciones y notificaciones, concursos y quiebras, otros	30172
Constitución S.A.	35408
Constitución S.A.S.	4710
Contrato S.R.L.	47270
Convocatorias	23162
Estatutos otras sociedades	146
Estatuto S.C.A.	6
Información y cultura	481
Modificaciones S.R.L.	26441
Partidos políticos	4538
Reforma otras sociedades	452
Reforma S.A.	35664
Remates Judiciales	8183
Sucesiones	103539

Transf. fondo de comercio 1994

5) Agrupación, estableciendo una base de referencia para el modelo de aprendizaje automáticos, procesando el set de entrenamiento normalizado (7.000 avisos) con AutoNLP, que lo agrupo y clasificó por tipos societarios, dividiendo el set en dos grupos (Entrenamiento, 70% y Test, 30%).
6) Creación y testeo de 7 modelos (Naïve Bayes NB, GS, Logistic regression LogR, linear regression LinR, random forest classifier RFC, K-Nearest Neighbor KNC, Support Vector Machines SVC) candidatos [8] en su respectiva notebook jupyter. Cada uno de ellos fue entrenado con el mismo 70% del set de 7.000 avisos y luego validado contra el 30% obtenido los resultados expresados en la (Tabla 2).

Tabla 2. Tabla comparativa de modelos de IA.

Nro	Model	Accuracy	f1	precision	recall
0	NB	86.90	86.56	88.04	86.90
1	GS	93.77	93.70	93.88	93.77
2	LogR	91.73	91.67	91.93	91.73
3	LinR	93.68	93.62	93.74	93.68
4	RFC	91.45	91.33	91.68	91.45
5	KNC	84.76	84.61	85.42	84.76
6	SVC	93.68	93.62	93.74	93.68

2 Creación de MVP

Este proyecto fue desarrollado mediante tecnologías agiles para la validación con los usuarios finales se crearon dos prototipos en base al diseño de dos productos mínimos viables (MVP) "*la versión de un nuevo producto que permite a un equipo recoger con el mínimo esfuerzo la máxima cantidad de conocimiento validado acerca de los consumidores"* [9], para luego realizar los testeo de usabilidad de los prototipos con él fin de corregir los errores de clasificación en la base de datos integrando el modelo IA.

1. MVP API - Client: Este prototipo consta de dos programas desarrollados con la librería flask para Python. Un API Rest que recibe el texto del aviso en formato JSON y retorna el código de clasificación y texto que identifica a la misma (ver Fig 1), de esta manera el usuario rápidamente a través de una aplicación web chequea si el tipo de aviso del sistema actual se corresponde con la clasificada por el modelo de IA.

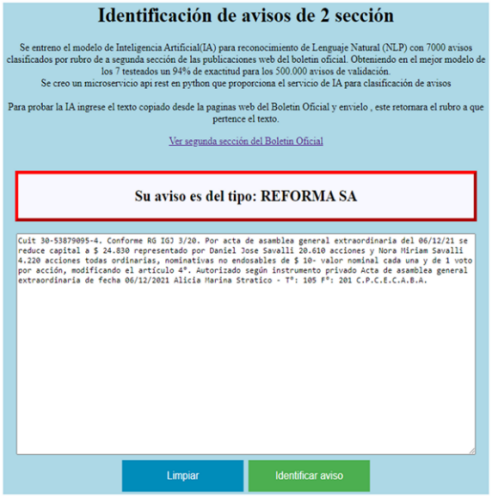

Fig. 1. Aplicación web y API clasificador de texto a tipo de aviso 2da sección, prototipo uno.

2. MVP Django: Este segundo prototipo (aplicación web) se desarrolló con Django con la posibilidad de utilizar diferentes modelos de IA (ver Fig. 2) generando una validación cruzada de los lotes de avisos que son seleccionados automáticamente por la aplicación desde una carpeta en disco con los archivos XML de los avisos. Una seleccionado el aviso a corregir se modifica el texto de mismo tanto en el archivo XML como también en la base de datos productiva.

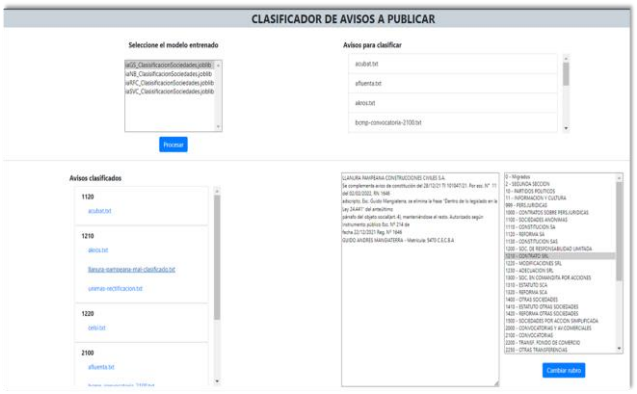

Fig. 2. Estructura general de la aplicación web Django en producción clasificador y modificador de avisos de 2da sección.

3 Líneas De Investigación Y Desarrollo

Los ejes principales del trabajo son:

- Implementar un modelo de IA que aprenda a identificar avisos de 3ra sección.
- Implementar un modelo que valide y permita depurar la base de datos existente.
- Evidencia la performance y vialidad del empleo de IA en el proceso de depuración de base de datos clasificadas.
- Crear dos MVP que permita validar su uso con los usuarios finales.

4 Resultados Obtenidos/Esperados

Una vez testear y comparar a los distintos modelos (Tabla 2) seleccionamos el de mayor eficacia aplicándolo a los 550.000 avisos disponibles en la 2da. Sección. Obteniendo una eficacia de entre el 84,76% y 93,77%.

Se obtuvo un total de 30.000 avisos sobre los 550.000 catalogados para revisión, de los cuales encontramos que sobre una muestra de 300 avisos 200 estaban mal clasificados y 100 no tenían el formato correspondiente o faltaban datos sobre el contenido del aviso.

Concluimos que ante las sucesivas migraciones de la base de datos a lo largo 20 años, se ha cambiado el formato y estructura de los avisos, pese a esto al ser un formato legal los principales indicadores están presentes permitiendo al modelo de aprendizaje supervisado aprender y clasificar con alto grado de asertividad (superiores al 84%).

El proceso total de trabajo desde la adquisición de los datos, depuración y entrenamiento demanda 3 días de trabajo, contra un proceso manual que se estima en más de un año si realizara de forma manual por personal de publicaciones con una dotación de 10 empleados.

Las características funcionales de los PMV definidas y ajustadas en 4 iteraciones con los usuarios finales, permitió estimar el tiempo de trabajo necesario para realizar la revisión y de ser necesario modificación de los avisos, con la alternativa de generar un nuevo set de datos de entrenamiento si se detectan avisos mal clasificados por la IA. La realización de una API permite también integrar rápidamente el sistema de clasificación por IA a los sistemas existentes tanto de escritorio como las aplicaciones web en el sitio del Boletín Oficial de la República Argentina que es de acceso público. La aplicación desarrollada con Django demostró ser muy ágil en la etapa de creación de la base de datos, validación de usuarios nativa conecta al active directory y definir un modelo escalable incremental en cada etapa de iteración de tecnología agiles. En cuanto a los modelos de clasificación por IA entrenados, se optaron por de mejor resultado GS, NB, RFC, SVC. Al realizar una validación cruzada de los modelos de IA encontramos que un 10% de los 30.000 avisos no tienen coincidencia plena (los 4 modelos clasifican de la misma forma). Para estos casos y dado que no encontramos diferencia desde la visión de programación se decidió implementar la opción de reclasificación para que los usuarios finales puedan generar un nuevo set de reentrenamiento en base a su criterio de clasificación con el fin de mejorar el modelo. El resultado es alentador pues una primera clasificación por la IA extrae los avisos dudosos en menos de 2 horas y en base a los MPV se estima que serían necesarias 4 semanas de trabajo para verificar y corregir las inconsistencias de los 30.000 avisos.

Como proyecto a futuro se contempla incorporar los modelos de IA a las aplicaciones existentes, tanto para validar el ingreso de los avisos, como para contar con una herramienta de validación histórica en tiempo real. Y aplicar el sistema de clasificación y depuración por IA a la 1ra y 3ra sección de la base de datos Boletín Oficial que supera los 2.000.000 de avisos.

También una nueva línea de investigación que se desprende del presente trabajo, sobre el análisis mediante inteligencia artificial de la base de datos legales. Como lo expresan varios autores "la clasificación automatizada de textos legales es un tema de investigación destacado en el campo legal. Sienta las bases para construir un sistema legal inteligente." [10]. Enmarcados dentro del proyecto de investigación y avances de los sistemas con inteligencia artificial en el sistema legal argentino y su aplicabilidad [11] y [12].

Otro punto a resaltar es que esta investigación, desarrollo de transferencia tecnológica se llevo a cabo con recursos propios en un termino de 6 meses, ahorrando al estado nacional varios miles de dólares, pues otros proyecto de implementación de sistemas tradicionales, para la depuración de base de datos han costado más del 100.000 dólares , mostrando no solo la factibilidad de realizar este proceso con IA, en un tiempo mucho menor, sino también el ahorro en utilizar trabajadores propios del organismo.

5 Formación De Recursos Humanos

El equipo se conformó por 2 docentes de maestría de la Universidad Abierta Interamericana (UAI), estudiantes del doctorado en informática UAI e investigadores en laboratorio de robótica física e inteligencia artificial (LRFIA) y un investigador colaborador. Este proyecto también cuenta con la participación de colaboradores externos del área de sistemas y publicaciones de boletín oficial. Se capacitó al equipo en técnicas de clasificación supervisada, tecnologías agiles, programación de sistemas web en Python y JavaScript, administración de base de datos PostgreSQL[13] y MSSQL como así también en machine learning empleando Google Colab[14] y Jupyter notebook [15].

6 Bibliografía

1. Chu, Xu & Ilyas, Ihab & Krishnan, Sanjay & Wang, Jiannan. Data Cleaning: Overview and Emerging Challenges. 2201-2206. 10.1145/2882903.2912574. https://dl.acm.org/doi/proceedings/10.1145/2882903 (2016).
2. Gudivada, Venkat & Apon, Amy & Ding, Junhua. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. International Journal on Advances in Software. 10. 1-20. https://www.researchgate.net/publication/318432363_Data_Quality_Considerations_for_Big_Data_and_Machine_Learning_Going_Beyond_Data_Cleaning_and_Transformations (2017).
3. Huxiao Liu, Lianhai Wang, Weinan Zhang, Wei Wang. An Illegal Billboard Advertisement Detection Framework Based on Machine Learning https://doi.org/10.1145/3358528.3358549 (2019).

4. Khayyat, Zuhair & Ilyas, Ihab & Jindal, Alekh & Madden, Samuel & Ouzzani, Mourad & Papotti, Paolo & Quiané-Ruiz, Jorge-Arnulfo & Tang, Nan & Yin, Si. (2015). BigDansing: A System for Big Data Cleansing. 10.1145/2723372.2747646. https://www.researchgate.net/publication/274256636_BigDansing_A_System_for_Big_Data_Cleansing
5. Badia Antonio. Data Science in the Database: Using SQL for Data Preparation. University of Louisville, USA - DOI: 10.4018/978-1-7998-9220-5.ch069 (2023).
6. Ley general de sociedades nº 19.550, t.o. 1984, http://servicios.infoleg.gob.ar/infolegInternet/anexos/25000-29999/25553/texact.htm
7. Paul Mooijman, Cagatay Catal, Bedir Tekinerdogan, Arjen Lommen, Marco Blokland. The effects of data balancing approaches: A case study,Applied Soft Computing, https://doi.org/10.1016/j.asoc.2022.109853.
8. Shovan Chowdhury, Marco P. Schoen, Research Paper Classification using Supervised Machine Learning Techniques - Intermountain Engineering, Technology and Computing (IETC) – IEEE https://ieeexplore.ieee.org/document/9249211 (2020).
9. Ries Eric Book The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses (2011).
10. Haihua Chen, Lei Wu, Jiangping Chen, Wei Lu, Junhua DingA comparative study of automated legal text classification using random forests and deep learning, Information Processing & Management,Volume 59, Issue 2, 2022, 102798, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2021.102798 (2022)..
11. Dobratinich Gonzalo. INTELIGENCIA ARTIFICIAL Y JUSTICIA: APLICABILIDAD DE LA TECNOLOGÍA EN LAS DECISIONES JUDICIALES EN ARGENTINA. Revista Direitos Culturais. https://doi.org/10.20912/rdc.v17i42.761
12. Gonzalo Ana Dobratinich (2021/09) - Evaluación De La Preparación Del Sistema Judicial Para La Adopción De Inteligencia Artificial – Universidad de San Andrés - http://hdl.handle.net/10908/18634 (2022/09/15), https://repositorio.udesa.edu.ar/jspui/handle/10908/18634
13. PostgreSQL “Documentation” https://www.postgresql.org/docs/ (2023)
14. Google Colab, “Documentation”, https://research.google.com/colaboratory/faq.html
15. Jupyter notebook “Juper Proyect Documentation”, https://docs.jupyter.org/en/latest/