

DEPURACIÓN DE BASES DE DATOS DE SEGUNDA SECCIÓN DEL BOLETÍN OFICIAL DE LA REPÚBLICA ARGENTINA MEDIANTE APRENDIZAJE DE MAQUINA

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

El Boletín Oficial de la República Argentina es una fuente importante de información para empresas, instituciones y particulares, pero la gran cantidad de información publicada diariamente hace que la depuración de bases de datos sea un proceso complejo y costoso. Se hipotetiza que el uso de un modelo de inteligencia artificial basado en aprendizaje de maquina (ML) puede aprender a clasificar y catalogar los avisos comerciales de manera eficiente y autónoma. La aplicación de técnicas de 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 y catalogación de los avisos comerciales de la segunda sección del Boletín Oficial de la República

Argentina, lo que podría representar un avance significativo en la depuración de bases de datos en este ámbito.

Palabras clave: Depuración, Base de datos, Aprendizaje de Maquina, Inteligencia Artificial, Aprendizaje Supervisado, Clasificación

CONTEXTO

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 (LRF). Las líneas de investigación sobre inteligencia dentro de los proyectos con financiamiento y duración a 2 años.

1. INTRODUCCIÓN

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 Machine Learning 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”* Chu [1]

Los problemas de calidad de los datos son complejos de resolver y existente gran cantidad de técnicas y herramientas para lograrlo Gudivada [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) Huxiao[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”* Khayyat [4].

Luego de realizar la (1) adquisición y depuración de datos se realizó el (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”* Badia [5]. 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. En el siguiente paso consistió en la (3) normalización de los datos eliminando los registros con

repubicaciones figura 1, pues el mismo aviso debe ser republicado varios días y con mismo texto. Para fines del entrenamiento solo necesitamos una sola muestra, reduciéndose el dataset a 550.000 registros.

idAviso	idRubro	RubroNombre	TextoXHTML	TextoXHTML_norm
0	89292	2300 AVISOS COMERCIALES	<Aviso>/> <NumeroTramite>#4344178# </NumeroTramite> % <NumeroPagDesde>11</NumeroPagDesde> %/in<Titulo>CENTRAL S.R.L./in</Titulo> <Texto>Por Reunión de Socios de fecha 27/03/2012, se resolvió por unanimidad aceptar la renuncia del Sr. Marcelo Eduardo José Costa a su cargo de gerente titular y designar al Sr. Marcos Pérez como gerente titular y al Sr. Gonzalo Jorge Mateos como gerente suplente. Por lo tanto, la gerencia queda integrada de la siguiente manera: Gerente Titular: Marcos Pérez; Gerente Suplente: Gonzalo Jorge Mateos. Los gerentes constituyen domicilio especial en Supacha 1111, Piso 18, Capital Federal; Luciana Zucatzosta autorizada por Acta de Reunión de Socios de fecha 27/03/2012./inAbogada - Luciana V. Zucatzosta/in</Texto> <Empieza>Vence>: 19/06/2012 N° 67509/12 v. 19/06/2012/in</Empieza> Vence>% <NumeroPagHasta>11</NumeroPagHasta> % </NumeroTramite>#4344178#</NumeroTramite> </Aviso>	central s.r.l por reunión socios fecha resolvió unanimidad aceptar renuncia sr marcelo eduardo jose costa cargo gerente titular designar sr marcos perez gerente titular sr gonzalo jorge mateos gerente suplente gerencia queda integrada siguiente manera gerente titular marcos perez gerente suplente gonzalo jorge mateos gerentes constituyen domicilio especial supacha piso capital federal luciana zucatzosta autorizada acta reunión socios fecha abogada luciana v zucatzosta

Figura 1: Registro base

Se procedió a la (4) agrupación por categorías de tipo de aviso figura 2, 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”* Mooijman[6].

RubroNombre	
AVISOS COMERCIALES	180641
BALANCES	527
CITACIONES Y NOTIFICACIONES, CONCURSOS Y QUIEBRAS, OTROS	30172
CONSTITUCION SA	35408
CONSTITUCION SAS	4710
CONTRATO SRL	47270
CONVOCATORIAS	23162
ESTATUTO OTRAS SOCIEDADES	146
ESTATUTO SCA	6
INFORMACION Y CULTURA	481
MODIFICACIONES SRL	26441
PARTIDOS POLITICOS	4538
REFORMA OTRAS SOCIEDADES	452
REFORMA SA	35664
REFORMA SCA	641
REINATES COMERCIALES	1248
REINATES JUDICIALES	8183
SUCESIONES	103539
TRANSF. FONDO DE COMERCIO	1994

Figura 2. Agrupación por categoría

Se procedió a establecer una (5) base de referencia con 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) Creamos 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 chowdhury[7] 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 figura 3.

	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	91.65	91.45
5	KNC	84.76	84.61	85.42	84.76
6	SVC	93.68	93.62	93.74	93.68

Figura 3. Resultados de modelos de IA

2. LINEAS 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.

3. RESULTADOS OBTENIDOS/ESPERADOS

Una vez probado en distintos modelos, se define el de mejor eficacia y se procede a aplicar 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.

El resultado es alentador y ya con el set de datos a verificar el personal de publicaciones estima un mes de trabajo para verificar y corregir las inconsistencias.

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 los modelos de aprendizaje a 1ra 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.”* Haihua [8]. Enmarcados dentro de los proyecto de investigación y avances de los sistemas con inteligencia artificial en el sistema legal Argentino y su aplicabilidad Dobratinich[9] y Dobratinich [10].

4. FORMACION DE RECURSOS HUMANOS

El equipo está formado por 1 docente Magister en tecnología informática graduado de la Universidad Abierta Interamericana (UAI), 1 alumnos de 5to año de ingeniería en sistemas (UAI) y un investigador del laboratorio de robótica física (LRF-UAI).

Este proyecto también cuenta con la participación de colaboradores externos del área de sistemas y publicaciones de boletín oficial.

5. BIBLIOGRAFIA

[1]Chu, Xu & Ilyas, Ihab & Krishnan, Sanjay & Wang, Jiannan. (2016). Data Cleaning: Overview and Emerging

Challenges. 2201-2206.
10.1145/2882903.2912574.

<https://dl.acm.org/doi/proceedings/10.1145/2882903>

[2] Gudivada, Venkat & Apon, Amy & Ding, Junhua. (2017). 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

[3] Huxiao Liu, Lianhai Wang, Weinan Zhang, Wei Wang (2019). An Illegal Billboard Advertisement Detection Framework Based on Machine Learning
<https://doi.org/10.1145/3358528.3358549>

[4] Khayyat, Zuhair & Ilyas, Ihab & Jindal, Alekh & Madden, Samuel & Ouzzani, Mourad & Papotti, Paolo & Quiané-Ruiz, Jorge-Arnulfo & Tang, Nan & Yin, Si. (2015). BigDancing: A System for Big Data Cleansing. 10.1145/2723372.2747646.
https://www.researchgate.net/publication/274256636_BigDancing_A_System_for_Big_Data_Cleansing

[5] Badia Antonio (2023). Data Science in the Database: Using SQL for Data Preparation . University of Louisville, USA - DOI: 10.4018/978-1-7998-9220-5.ch069

[6] 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>.

[7] Shovan Chowdhury, Marco P. Schoen (2020), Research Paper Classification using Supervised Machine Learning Techniques - Intermountain Engineering, Technology and Computing (IETC) – IEEE
<https://ieeexplore.ieee.org/document/9249211>

[8] Haihua Chen, Lei Wu, Jiangping Chen, Wei Lu, Junhua Ding (2022). A 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>.

[9] Dobratinich Gonzalo (2022). INTELIGENCIA ARTIFICIAL Y JUSTICIA: APLICABILIDAD DE LA TECNOLOGÍA EN LAS DECISIONES JUDICIALES EN ARGENTINA. Revista Direitos Culturais. 17. 203-216. 10.20912/rdc.v17i42.761.

https://www.researchgate.net/publication/363593342_INTELIGENCIA_ARTIFICIAL_Y_JUSTICIA_APLICABILIDAD_DE_LA_TECNOLOGIA_EN_LAS_DECISIONES_JUDICIALES_EN_ARGENTINA

[10] Gonzalo Ana Dobratinich (2021)- Evaluación De La Preparación Del Sistema Judicial Para La Adopción De Inteligencia Artificial – Universidad de San Andrés -
<https://repositorio.udes.edu.ar/jspui/handle/10908/18634>