Agentes y Sistemas Inteligentes

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

AUTORES:

Néstor A. Balich nestor.balich@uai.edu.ar
Franco A. Balich francoadrian.balich@uai.edu.ar
Hugo Fraga hfraga@boletinoficial.gob.ar
Filiación: CAETI - Centro de Altos Estudios en Tecnología Informática
Universidad Abierta Interamericana (UAI)
Laboratorio de robótica física e inteligencia artificial

Ingeniería en Sistemas Informáticos

LÍNEAS DE INVESTIGACIÓN:

Automatización y Robótica

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.

INTRODUCCIÓN:

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. 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]

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.
- Evidenciar la performance y vialidad del empleo de IA en el proceso de depuración de base de datos clasificadas

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 validad 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].

	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

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
REMATES COMERCIALES	1248
REMATES JUDICIALES	8183
SUCESIONES	103539
TRANSF. FONDO DE COMERCIO	1994

	idAviso	IdRubro	RubroNombre	TextoXHTML	TextoXHTML_norm
0	89292	2300	AVISOS COMERCIALES	<Aviso>\r\n <NumeroTramite>#F4344178I# </NumeroTramite> % <NumeroPagDesde>11</NumeroPagDesde> %\r\n<Titulo>CENTRAPAL S.R.L\r\n</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 Pérés 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 Pérés; Gerente Suplente: Gonzalo Jorge Mateos. Los gerentes constituyen domicilio especial en Suipacha 1111, Piso 18, Capital Federal. Luciana Zuccatosta autorizada por Acta de Reunión de Socios de fecha 27/03/2012.\r\nAbogada – Luciana V. Zuccatosta\r\n</Texto><Empieza-Vence>e. 19/06/2012 Nº 67509/12 v. 19/06/2012\r\n</Empieza-Vence>% <NumeroPagHasta>11</NumeroPagHasta> % <NumeroTramite>#F4344178F#</NumeroTramite></Aviso>	centrapal srlpor reunión socios fecha resolvió unanimidad aceptar renuncia sr marcelo eduardo josé costa cargo gerente titular designar sr marcos pérs gerente titular sr gonzalo jorge mateos gerente suplente gerencia queda integrada siguiente manera gerente titular marcos pérs gerente suplente gonzalo jorge mateos gerentes constituyen domicilio especial suipacha piso capital federal luciana zucatosta autorizada acta reunión socios fecha abogada luciana v zucatosta

time: 74.1 ms (started: 2022-02-03 19:19:50 +00:00)

Art: 14421

REFERENCIAS:

- Gonzalo Ana Dobratinich (2021)- Evaluación De La Preparación Del Sistema Judicial Para La Adopción
- De Inteligencia Artificial – Universidad de San Andrés Dobratinich Gonzalo (2022). INTELIGENCIA ARTIFICIAL Y JUSTICIA: APLICABILIDAD DE LA TECNOLOGÍA EN LAS DECISIONES JUDICIALES EN ARGENTINA. Revista Direitos Culturais.
- 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 & ManagementShovan
- Chowdhury, Marco P. Schoen (2020), Research Paper Classification using Supervised Machine Learning Techniques - Intermountain Engineering, Technology and Computing (IETC)
- Paul Mooijman, Cagatay Catal, Bedir Tekinerdogan, Arjen Lommen, Marco Blokland. The effects of data balancing approaches: A case study,Applied Soft ComputingBadia Antonio (2023). Data Science in the Database: Using SQL for Data Preparation . University of Louisville, USA
- 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
- Huxiao Liu, Lianhai Wang, Weinan Zhang, Wei Wang (2019). An Illegal Billboard Advertisement Detection Framework Based on Machine Learning