



MAESTRÍA EN EXPLOTACIÓN DE DATOS Y GESTIÓN DEL
CONOCIMIENTO

Proyecto de Tesis

Clasificación de los dígitos escritos en los telegramas de las elecciones
legislativas en Santa Fe mediante técnicas de adaptación de dominio.

Autor: Franco Lianza

Tutor: Dr. Leandro Bugnon

Julio 2022

Resumen

En Argentina se celebran elecciones cada 2 años, lo cual implica que se debe incurrir en altos costos para poder llevarlas a cabo. Al finalizar el día del sufragio, cada una de las mesas de voto arman una planilla con los votos obtenidos por cada partido político. La misma es enviada al correo en forma de telegrama para luego ser escaneado y enviado a un centro de cómputo. El proceso por el cual se digitalizan a mano cada uno de los telegramas los telegramas en un sistema central es costoso, ineficiente y susceptible a errores humanos. Contar con una herramienta que permita digitalizar automáticamente cada uno de los telegramas al sistema central de cómputo, agilizará y aportará transparencia al proceso reduciendo los costos y errores. Por tal motivo, se centrará en desarrollar un proceso de extracción de dígitos de los telegramas para luego entrenar distintos modelos que sean capaces de clasificarlos. De esta manera, se podrán definir métricas que permitan seleccionar el mejor de ellos y disponibilizarlo en las próximas elecciones.

Motivación e importancia del campo

Existen, principalmente, tres tipos de elecciones:

- Elecciones nacionales, para elegir a las autoridades federales del país: el Poder Ejecutivo, constituido por el Presidente y el vicepresidente y el Congreso Nacional, formado por Senadores y Diputados.
- Elecciones provinciales y de la Ciudad de Buenos Aires o locales, para elegir a las autoridades de cada provincia: los poderes ejecutivos de las provincias y sus legislaturas.
- Elecciones municipales, regidas por las leyes y procedimientos de cada provincia.

Si bien emitir el sufragio es diferente en cada una de ellas, generalmente consta de ingresar a un cuarto oscuro, elegir el candidato que se desea y depositar el voto en una urna. Al finalizar la jornada, las autoridades de mesas recuentan los votos y llenan una planilla a mano alzada donde se resume la cantidad de votos obtenidos por cada candidato o partido político. Dicha planilla es escaneada y enviada a través de un telegrama del correo argentino al centro de cómputo para su procesamiento. Una vez allí, se contabilizan en un sistema informático a partir de un grupo de empleados. Finalmente, se procede a una etapa de validación de los datos para constatar que lo computado coincide con lo escrito.

Para que el proceso sea lo más rápido y eficaz posible, se requiere a una gran cantidad de personas destinadas al centro de cómputo. Tal solución hace que el proceso sea altamente ineficiente en cuanto a tiempos y costos se refiere. En las elecciones legislativas del 2021 se gastaron unos \$17.000 millones de pesos de los cuales \$4.000 millones de pesos fueron destinados a sueldos para el personal¹.

Requerimientos y desafíos

La clasificación de números es un problema que, si bien parece resuelto con [?] y la creación del dataset MNIST, no debe ser tomada a la ligera. Si bien la rama de *Computer Vision*, que se encarga de desarrollar herramientas capaces de reconocer patrones complejos en imágenes, se encuentra en un punto en el cual presenta resultados óptimos en la mayoría de los casos [?, ?]; no se encuentra exenta al problema del sesgo en los datos.

No existe una única forma de escribir y año a año cambian las personas que son los jefes de mesa encargados de completar los telegramas. Las características de los números escritos difiere entre cada elección. Cuando la distribución de los datos de entrenamiento difiere a la de los datos de aplicación, se está ante un *corrimiento de dominio* (*domain shift* o *data drift* en inglés). Esto provoca los siguientes problemas:

- No se puede utilizar un modelo entrenado en telegramas de elecciones pasadas.
- Para entrenar algún modelo de clasificación de imágenes se requiere de una gran cantidad de datos y sus respectivas etiquetas.

¹Fuente: [El cronista](#)

Es por lo expuesto que no puede utilizarse un esquema de entrenamiento convencional. Una alternativa es la realizada en trabajos anteriores, donde se aplicaron distorsiones al conjunto de entrenamiento para aumentar la cantidad de datos de entrenamiento y que de esta forma el modelo pueda generalizar y aplicarse a los telegramas de elecciones de la Ciudad de Buenos Aires [?]. No obstante, no se muestran métricas concretas respecto de los resultados obtenidos en los telegramas reales.

Problemas no resueltos

El problema que quiere resolver es la digitalización automática de los telegramas mediante machine learning de forma robusta.

La digitalización de los telegramas de las elecciones sigue siendo de forma manual, lo que genera lentitud y desconfianza en el proceso. Detectar los votos de manera automática y reducir la intervención humana, aumentará la eficiencia de las elecciones y la confianza en ellas.

Solución propuesta

Durante el proceso de digitalización de los telegramas en una elección, se ejecutará el proceso que se encarga de extraer los dígitos de los mismos y entrenará un modelo de clasificación específico a partir de *transfer learning*. Luego, se aplicará el modelo y su salida será el voto digitalizado junto a su probabilidad. De esta manera, la digitalización estará realizada y se podrá proceder al proceso de validación humana de las predicciones.

El *transfer learning* (transferencia de aprendizaje) es un área del *machine learning* que se encarga de almacenar el conocimiento ganado en un problema y aplicarlo a otro [?]. Sin embargo, como no se dispondrán de las etiquetas al momento de la elección, se utilizarán técnicas de *adaptación de dominio* (AD). La misma consiste en la habilidad de aplicar un modelo entrenado en uno o mas dominios de origen (*source domain*) en un uno distinto pero relacionado (*target domain*) [?]. A modo de ejemplo, la figura ?? muestra dos conjuntos de datos de dígitos pero con dominios diferentes.

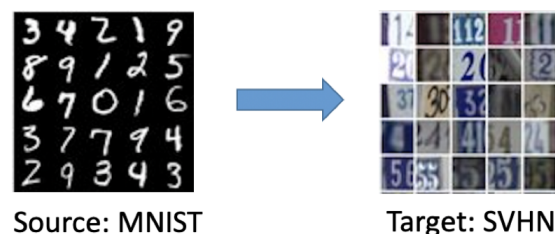


Figura 1: Ejemplo dominios diferentes: MNIST y SVHN

Objetivo del trabajo

Como objetivo general del trabajo se pretende desarrollar un proceso que permita transformar los telegramas de elecciones legislativas de Santa Fe, entrenar un modelo de clasificación mediante una técnica de adaptación de dominio y poder digitalizar los telegramas.

Esto desprende los siguientes objetivos específicos:

- Armar el proceso de ETL de los telegramas que permita limpiar y extraer los dígitos.
- Determinar un conjunto de datos etiquetado a ser utilizado como dominio de origen.
- Entrenar distintas arquitecturas de redes convolucionales mediante técnicas de adaptación de dominio.
- Analizar y evaluar las métricas que permitan seleccionar el mejor modelo.

- Seleccionar el mejor par modelo - técnica AD.

Metodología

Se hará revisión del estado del arte para después continuar con el preprocesamiento de los datos. Los telegramas serán descargados desde la [página oficial del estado argentino](#). Luego se extraerán los dígitos de los votos de cada telegrama utilizando la librería *OpenCV* [?] y Python. Se realizarán ciclos de entrenamiento de distintas redes convolucionales (LeNet [?], ResNet [?], etc) en el dominio de origen para luego aplicar técnicas de AD que permita transferir el conocimiento al dominio de destino (telegramas legislativos de Santa Fe). Se evaluarán las métricas de cada uno de ellos y se procederá a elegir el mejor.

Plan de Trabajo

El trabajo se realizará a partir de las siguientes etapas:

- Estudio del estado del arte: estudiar las diferentes técnicas de *domain adaptation*.
- Extracción datos: recolectar los telegramas y extraer los dígitos de los votos.
- Limpieza de datos: detectar votos incorrectos (por ejemplo letras) y tratarlos para reducir lo máximo posible el ruido en los datos.
- Experimentación: realizar ciclos de entrenamiento de modelos y adaptarlos mediante alguna de las técnicas disponibles.
- Elaboración de reportes: generar reportes que sinteticen los experimentos realizados para compararlos.
- Redacción de tesis.

Se estima que el proyecto comienza en Agosto del 2022 y finaliza en Diciembre del 2022. La figura ?? presenta el cronograma propuesto.

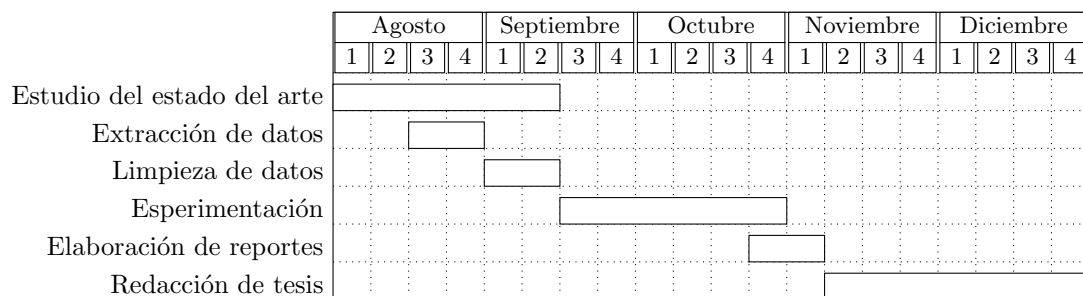


Figura 2: Distribución de las tareas del proyecto.

Bibliografía

[2]