UNIVERSIDAD DE LA RIOJA

TRABAJO FIN DE ESTUDIOS

Título

Fact checking sobre visualizaciones de datos

Autor/es

MARINA BLANCO IGLESIAS

Director/es

ÁNGEL LUIS RUBIO GARCÍA

Facultad

Escuela de Máster y Doctorado de la Universidad de La Rioja

Titulación

Máster universitario en Ciencia de Datos y Aprendizaje Automático

Departamento

MATEMÁTICAS Y COMPUTACIÓN

Curso académico

2021-22

publicaciones.unirioja.es
E-mail: publicaciones@unirioja.es

Trabajo de Fin de Máster

Fact checking sobre visualizaciones de datos

Fact checking on data visualizations

Autora : Blanco Iglesias, Marina

Tutor: Rubio García, Ángel Luis

MÁSTER:
Ciencia de Datos y Aprendizaje Automático

Escuela de Máster y Doctorado

UNIVERSIDAD DE LA RIOJA

AÑO ACADÉMICO: 2021/2022

Resumen

En un contexto marcado por la continua difusión de informaciones falsas (fake news), las iniciativas dedicadas a la verificación de datos (fact checking) se han convertido en herramientas esenciales para comprobar y revisar los datos y la información. Dado que las gráficas han resultado ser contenidos cada vez más consumidos y se ha incrementado su relevancia como formato informativo, parece oportuno hacer un estudio, examinar y analizar la posición que el fact checking ocupa en el proceso de visualización de datos.

Tras este análisis se han observado las carencias y limitaciones existentes en el ámbito de la visualización de datos y se han aplicado técnicas de la Ciencia de Datos para plantear un posible primer acercamiento para solventar esta falta de recursos.

Abstract

In a context marked by the continuous spreading and transferring of false, inaccurate and misleading information(fake news), initiatives dedicated to the task of fact checking have become essential tools in testing and reviewing data.

Since graphics have turned out to be increasingly consumed contents and their relevance as an informative format has risen significantly, it seems appropiate to carry out a research, study and analysis of the place that fact checking occupies within data visualization process.

Over the analysis period, the existing deficiencies and limitations in the field of graphical presentation of data have been observed and Data Science Techniques have been applied to propose a posible first approach to solve this lack of resources and solutions designed and developed in this field.

Índice

Introducción

La nueva era digital ha cambiado por completo la forma en la que se transmite la información en el día a día. Internet, el desarrollo de los medios digitales, las redes sociales y las nuevas tecnologías han facilitado el intercambio de información sin fronteras. La accesibilidad, inmediatez, conectividad, la interacción en tiempo real, la viralidad… son características que forman parte de nuestro entorno digital y a las que todos estamos habituados. Cualquier persona, desde cualquier lugar, es capaz de difundir una información que casi al momento pueden recibir miles o millones de personas. Esta inmediatez para recibir las informaciones, así como la facilidad para publicarlas y la accesibilidad por parte de todos han traído consigo un serio inconveniente que es la falta de tiempo para verificar y contrastar la ingente cantidad de noticias y datos que nos llegan constantemente.

Esta sobreabundancia y exceso de datos e información sin contrastar ha generado un importante nivel de desinformación, de confusión en la información y de proliferación de datos incorrectos o directamente falsos. Esto ha provocado la aparición de las "fake news" (noticias falsas), que han entrado a formar parte de nuestro día a día. De esta forma, lo que se prometía como la panacea de la comunicación conlleva otros aspectos no tan positivos y a los que se intenta poner freno. Es precisamente por la proliferación de las "fake news" y la desinformación generalizada por la saturación de datos e informaciones, por lo que se empieza a tener conciencia de la importancia de la información veraz y de calidad, y la relevancia de atajar el problema y buscar herramientas para combatir esta tergiversación. En este punto en el que se buscan soluciones aparece el "fact checking" (verificación de datos), como un proceso para verificar las informaciones, luchar contra las "fake news", promover la transparencia y determinar la veracidad y corrección de la información y los datos.

Este proyecto está dedicado al "fact checking" sobre visualización de datos. La importancia de las tablas y los gráficos es cada vez mayor en nuestros días, ya que permiten condensar la información en unidades visuales de consumo rápido. Por tanto, la información recogida en las visualizaciones de datos es especialmente sensible a incluir errores, falsedades, ausencias de rigor, mentiras e inexactitudes.

Por otra parte, la Ciencia de Datos es un campo interdisciplinar que implica el uso de métodos científicos, procesos y sistemas para un mejor entendimiento de los datos y para extraer información y conocimiento en la toma de decisiones. En la Ciencia de Datos confluyen técnicas y teorías extraídas de las matemáticas, la estadística y la informática.[1] En particular, la visualización de datos es una rama de la Ciencia de Datos que nos da pautas acerca de cómo representar datos en formato gráfico de forma

[1] Definición de la Universidad de La Rioja en su descripción del máster https://www.unirioja.es/estudios/master/855M/index.shtml

correcta. De hecho, la mayoría de los proyectos de ciencia de datos representan sus resultados mediante visualizaciones de datos por medio de gráficos.

Si unimos los dos mundos descritos en esta introducción, poder aplicar los conocimientos combinados que aporta la Ciencia de Datos (preparación de datos, minería de datos, análisis predictivo y aprendizaje automático, entre otros) supone una interesante oportunidad para sondear las posibilidades que estos conocimientos nos pueden brindar en el proceso de “fact checking”, centrándonos en particular en el ámbito de las visualizaciones de datos.

Análisis

¿Qué es el fact-checking y qué relación tiene con las visualizaciones de datos?

El "fact checking" es el proceso de verificar la información mediante diferentes métodos, pero con un mismo objetivo: luchar contra las noticias falsas utilizando la inteligencia artificial junto con la verificación humana.

La visualización de datos es la representación gráfica de la información y los datos, mediante elementos visuales como gráficas, diagramas, mapas y tablas. Esta representación de datos de forma visual es una manera accesible de detectar y comprender las tendencias, los valores atípicos y los patrones en los datos. Se trata de una herramienta relativamente fácil de utilizar y que se puede aplicar en todos los ámbitos. Con la visualización de datos exploramos, comprendemos y podemos tomar decisiones, así como optimizar el tiempo y los procesos. Para ello la visualización de datos debe plasmar información útil y que esa información quede representada de forma correcta y clara. En concreto, el objetivo de una buena gráfica debe ser representar los datos de forma adecuada y con la mayor rigurosidad posible para poder comprender fácilmente e interpretar dichos datos. Las gráficas, de por sí, tienden a generar una falsa sensación de veracidad debido a nuestro bagaje cultural y nuestro aprendizaje. Toda información que nos llega parece más veraz si viene respaldada por unos datos numéricos (los números nos transmiten objetividad y rigurosidad) y por esta sensación de exactitud y veracidad las gráficas pueden conseguir el efecto contrario.

Los datos al igual que la información pueden ser objetivos y reales pero la forma de plasmarlos puede estar sujeta a la subjetividad o la distorsión de la realidad de quien nos lo está transmitiendo. Las gráficas no siempre son rigurosas, ni dicen la verdad. De hecho, las gráficas mienten y más de lo que creemos.

¿Por qué es necesario el fact-checking en las visualizaciones de datos?

Cualquier visualización de datos, independientemente de la rigurosidad con la que se haya plasmado, nos puede llevar a engaño si no prestamos la suficiente atención. ¿Quién no ha escuchado alguna vez y ha hecho suya la expresión: "una imagen vale más que mil palabras"? Esta expresión cobra un sentido especial en el contexto de este bombardeo mediático del que todos formamos parte, y a dicha frase tendríamos que añadir: "…siempre que dispongas del contexto y la formación necesaria para leer e interpretar esa imagen de forma correcta". Los números y las gráficas pueden resultar muy persuasivos porque van asociados a la ciencia, a la razón, dando una falsa sensación de objetividad y precisión, llegando a ser muy convincentes. De ahí que la formación, la tecnología y la verificación son factores que se deben tener en cuenta para ganar la

batalla a la desinformación y manipulación. Es necesario que aprendamos a leer, comprender e interpretar la plasmación de datos e información por medio de las gráficas. Una gráfica no es una mera ilustración que acompaña un texto y no debemos contemplarla como una simple fotografía o dibujo. Nos preguntamos ¿qué es un gráfico? Según la definición de la RAE[2] un gráfico/a es:

> *Del lat. graphĭcus, y este del gr. γραφικός graphikós.*
>
> *1. adj. Perteneciente o relativo a la escritura y a la imprenta.*
>
> *2. adj. Dicho de una descripción, de una operación o de una demostración: Que se representa por medio de figuras o signos. U. t. c. s.*
>
> *3. adj. Dicho de un modo de hablar: Que expone las cosas con la misma claridad que si estuvieran dibujadas.*
>
> *4. m. Representación de datos numéricos por medio de una o varias líneas que hacen visible la relación que esos datos guardan entre sí.*
>
> *5. f. gráfico (|| representación por medio de líneas).*

A excepción de la primera entrada, el resto de las acepciones nos aclaran lo que las gráficas son y deberían plasmar. Una gráfica nos debe aportar claridad.

Como enumera Dürsteler[3] el objetivo de una gráfica es:

- Comunicar un mensaje.
- Mostrar grandes cantidades de información de una forma que sea fácil de comprender y compacta.
- Revelar los datos (Qué ocurre, causa-efecto).
- Controlar la evolución de diferentes parámetros de forma periódica.

Al final, la idea fundamental es contemplar una gráfica como un acto comunicativo: siempre nos va a querer transmitir un mensaje. Al igual que todo proceso comunicativo las gráficas tienen: un contexto, un emisor, un receptor, un mensaje y un canal para transmitir ese mensaje. Si no se tiene en cuenta alguno de los elementos que forman parte del acto de la comunicación estaremos ante un proceso fallido o no eficaz. Las gráficas poseen su propio lenguaje visual, vocabulario de símbolos, una gramática y un contexto. Existe una analogía entre las gráficas y los conceptos lingüísticos[4]: una gráfica corresponde a la oración, una infografía o grupo de gráficas es el párrafo, la magnitud representada es el sujeto de la oración, la impresión visual sería el predicado, una gráfica con una serie de datos correspondería a una oración simple, mientras que si tienen más de una serie de datos se corresponden con una oración compuesta, y si las gráficas están relacionadas con otras en una infografía hablaríamos de una oración subordinada.

[2] RAE: Real Academia Española

[3] Libro "Visualización de información" de Juan Carlos Dürsteler

[4] Analogía de Joaquín Sevilla Moróder en su libro "Gramática de las gráficas"

Como ya hemos visto, las gráficas son un medio de comunicación y siguiendo con la analogía lingüística deben tener un ortografía, sintaxis, semántica y estilos literarios. Por tanto, para entender bien una gráfica debes comprender su vocabulario y su sintaxis. De la misma forma que cuando leemos un texto o un artículo debemos hacerlo hasta el final para entenderlo (y no quedarnos con el título), lo mismo debemos hacer con las gráficas: si queremos entenderlas necesitamos tomarnos un tiempo porque no es lo mismo ver una gráfica que entenderla. Una gráfica será adecuada cuando conlleve un mensaje que trasmitir, dicho mensaje esté soportado por los datos y la gráfica lo trasmita adecuadamente.

Según Alberto Cairo en su libro "How Charts Lie" las gráficas no siempre son exactas y pueden llevarnos a error por diferentes motivos:

- Pobreza en el diseño.
- Uso de datos erróneos.
- Presentación de una cantidad de datos inapropiada, ya sea por ser escasa o excesiva
- Ocultación o confusión por incertidumbre.
- Sugerencia de patrones engañosos.
- Refuerzo de nuestras necesidades o prejuicios.

Para ilustrar adecuadamente estas situaciones, a continuación mostraré algunas gráficas recientes, en las que se aprecian los diferentes aspectos por los que una gráfica miente o no se ajusta a la realidad.

POBREZA EN EL DISEÑO

Son muchos los aspectos que hay que tener en cuenta a la hora de representar datos en gráficas. Las distorsiones visuales son fáciles de identificar para aquellos que saben leer bien una gráfica, pero pueden sesgar nuestras percepciones si no se les presta la suficiente atención. El tamaño de los símbolos puede no ser proporcional a los datos, la escala puede no estar bien seleccionada o buscar trasmitir una idea diferente a la que se representa añadiendo efectos 3D. El diseño de una gráfica, sus escalas y codificaciones, deben depender de la naturaleza de los datos. Una buena gráfica te debería permitir visualizar patrones y tendencias sin necesidad de leer todos los números.

En la siguiente gráfica podemos apreciar como la elección del tipo de gráfica no ha sido la más acertada para reflejar los datos que se pretendían. Percibimos que la similitud en el tamaño de los sectores no nos deja ver a simple vista el dato cuantitativo y tenemos que leer los porcentajes para saber qué empresas manufacturan más. Si se hubiesen plasmado los datos por ejemplo en una gráfica de barras apreciaríamos con mayor facilidad el mensaje que se pretende trasmitir.

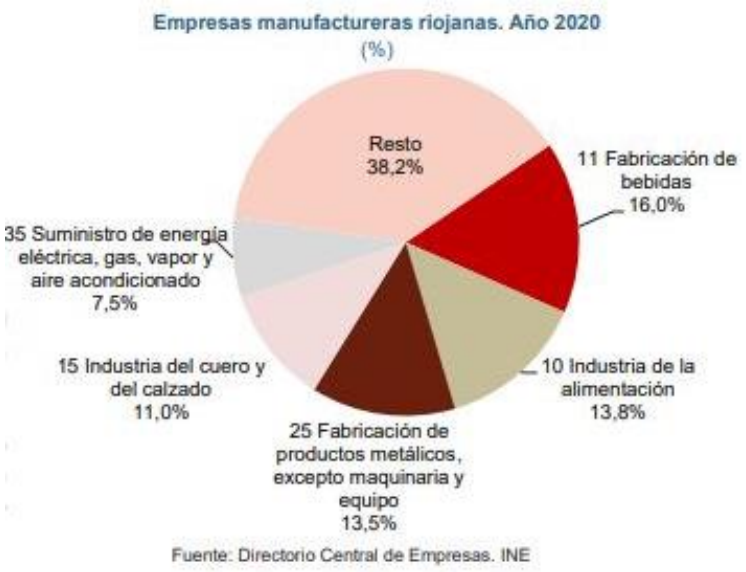


IMAGEN 1 GRÁFICA DEL INSTITUTO DE ESTADÍSTICA DE LA RIOJA

La siguiente gráfica contiene varios errores que debemos evitar a la hora de realizar una. En primer lugar, observamos que la ubicación de la referencia (0) se encuentra en un valor menor que 0 (-0.8). Otro aspecto que contemplamos en el eje Y son los intervalos irregulares y que algunos valores negativos han sido intercambiados.

IMAGEN 2 GRÁFICA DE UN COMUNICADO OFICIAL DEL GOBIERNO DE URUGUAY DE FECHA 03/02/2022

Otro ejemplo de pobreza en el diseño se da cuando la propia gráfica nos tiene que dar una guía de cómo entenderla.

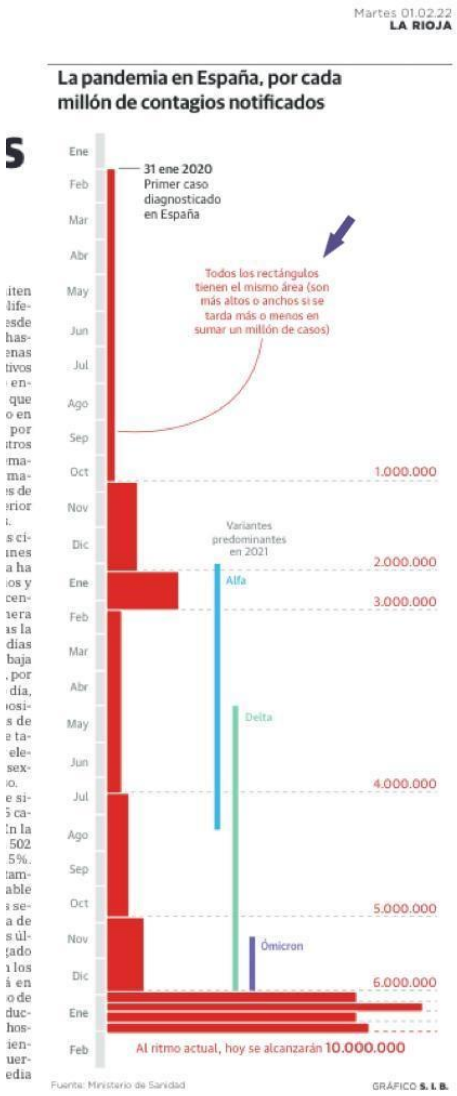


IMAGEN 3 GRÁFICA DEL PERIÓDICO "LA RIOJA" DE FECHA 01/02/2022

Las gráficas en 3D constituyen un caso aparte: están por todos lados, son ostentosas e impresionantes, pero suelen fracasar a la hora de informarnos. Eligiendo el ángulo adecuado se puede cambiar la perspectiva y resaltar lo que le interesa al emisor. Si estos gráficos en 3D fuesen interactivos o vistos a través de realidad virtual el receptor podría moverse alrededor del gráfico para verlo en dos dimensiones, lo que cambiaría nuestra percepción. Al igual que tienen detractores, sus defensores sostienen que los efectos 3D son acertados y que siempre se puede escribir cada dato que el gráfico muestra en la parte superior de las barras, líneas o segmentos. Entonces, ¿para qué diseñar la gráfica en primer plano?

La siguiente gráfica es un ejemplo claro de que el efecto 3D puede distorsionar la realidad de los datos. Observamos cómo según la perspectiva que se ha tomado, apreciamos en primer plano el segmento del turismo rural, y este segmento da la sensación de ser el más grande de los tres. Sin embargo, los datos numéricos nos muestran que estamos totalmente equivocados.

IMAGEN 4 GRÁFICA INSTITUTO DE ESTADÍSTICA DE LA RIOJA

El siguiente gráfico muestra la precipitación acumulada en diferentes años con un gráfico 3D. Se intenta ser efectista con la gráfica utilizando un efecto de "depósito de agua" con una imagen en 3D para representar precisamente la precipitación acumulada. Pero la gráfica consigue el efecto contrario puesto que en ambas el dato queda desdibujado dentro del propio efecto 3D de la gráfica. Además, se echa de menos una escala en el eje X y la que aparece en el Y es errónea porque el valor 0 no está situado sobre la base y no tiene sentido que las precipitaciones tengan valores negativos.

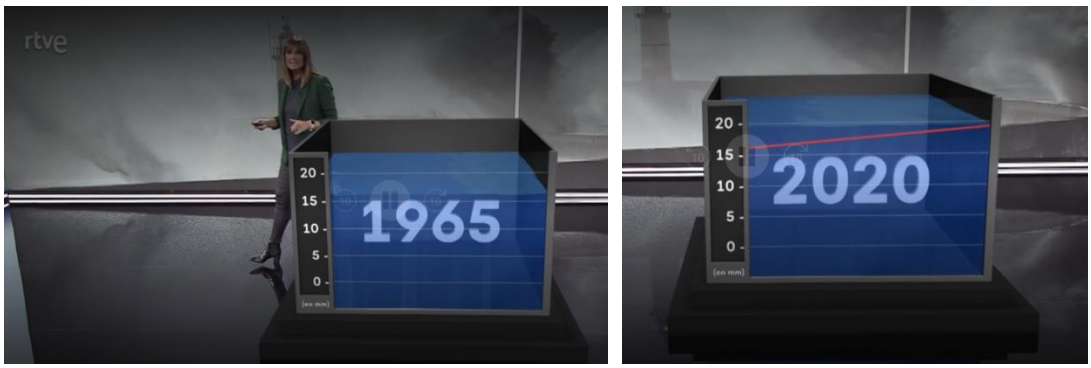


IMAGEN 5 GRÁFICAS DEL DEL TELEDIARIO RTVE 08/11/2021

USO DE DATOS ERRÓNEOS O POCO FIABLES

Si atendemos a la expresión "Garbage in, garbage out", muy extendida y usada entre los informáticos, su sentido metafórico se ajusta y puede aplicarse al contexto de las gráficas. "Basura entra, basura sale" (que es su significado literal), hace referencia a la

calidad de los productos y las informaciones. Si la base o el fundamento de un producto es malo o erróneo, el resultado generalmente lo será. Haciendo una similitud con las gráficas, éstas pueden tener una apariencia bonita, sorprendente, interesante o curiosa, pero si codifican datos erróneos ya partimos de una base fallida (la basura que aparece en la expresión, por supuesto hablando en sentido figurado) que nos va a confundir y ese gráfico no será válido en tanto en cuanto no se ajusta a la realidad desde su origen. En la actualidad no podemos esperar que todo el mundo sea capaz de verificar la exactitud de los datos que diariamente nos muestran las gráficas. La falta de tiempo y de conocimiento son aspectos claves que nos llevan a hacer un ejercicio de confianza. Como pautas básicas no debemos confiar en gráficas cuya fuente no nos es familiar hasta que no comprobemos la fuente. También debemos desconfiar de autores o publicaciones que no mencionan la fuente de sus datos o que no nos redirigen a dicha fuente. En ocasiones la ignorancia o falta de conocimiento es la causa de una mala gráfica, o la no profesionalidad en la materia, la adopción de posiciones partidistas o tendenciosas, o simplemente el descuido. Hoy más que nunca hay que ser críticos cuando observamos una gráfica.

Un ejemplo muy claro de este problema que se repite en la actualidad se da cuando consultamos los datos del coronavirus. Es fácil encontrar una gráfica que represente la situación de la pandemia, pero resulta difícil encontrar dos gráficas que reflejen los mismos datos correspondientes al mismo día. Los siguientes gráficos son un ejemplo de esta disparidad de datos. Si tomamos un par de fechas determinadas, tales como el 17 de enero y el día siguiente, vemos que los datos son totalmente diferentes. En el primer gráfico el día 17 registra 5305 nuevos casos, al día siguiente 1061 y en el otro gráfico 775 y 739 respectivamente. Como podemos ver los datos son totalmente dispares cuando se supone que hacen referencia al mismo día y al mismo lugar.

IMAGEN 6 GRÁFICA DEL BUSCADOR DE GOOGLE

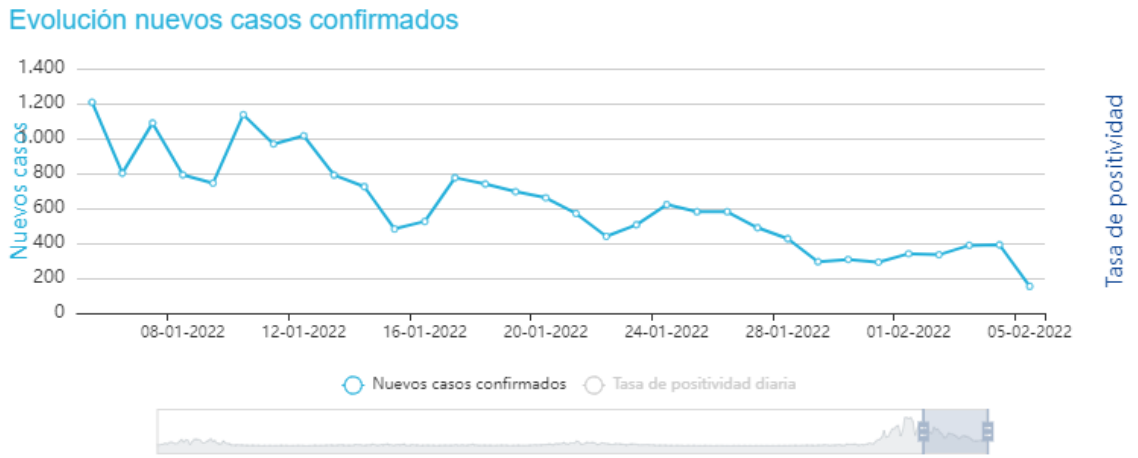


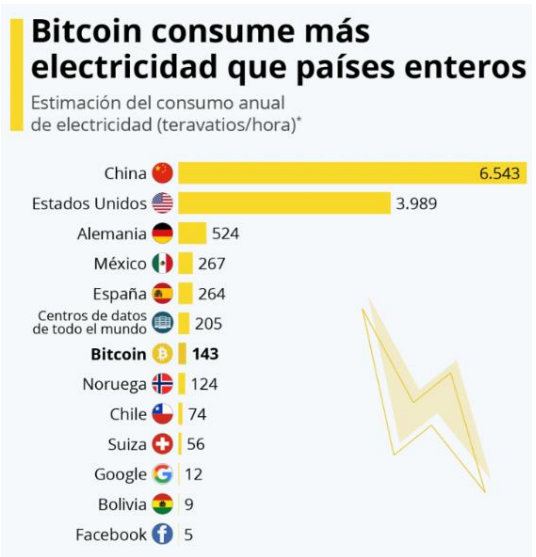
IMAGEN 7 GRÁFICA DE LA PLATAFORMA DE BI OFICIAL DE LA RIOJA

PRESENTACIÓN DE UNA CANTIDAD INAPROPIADA DE DATOS

Otra forma de confundir es la selección selectiva de datos ("cherry- picking data") que se van a visualizar. Podemos hacer resaltar un dato determinado entre otros muchos de dos maneras diferentes. Por un lado, podemos remarcar el dato que nos interesa eligiendo cuidadosamente los datos que lo acompañan y obviar de forma malintencionada los datos que puedan refutar la información que nos interesa trasmitir. Por otro lado, otra alternativa sería hacer lo contrario, visualizar la mayor cantidad posible de datos en el gráfico para distorsionar o desbordar la percepción de aquellos a los que quieres persuadir. Se trata de desfigurar la realidad que muestran los datos y que no nos interesa resaltar o meramente solapar esos datos.

Si nos fijamos en los siguientes ejemplos, podemos ver claramente como el mismo dato ha sido representado de dos formas muy diferentes y visualmente nos transmiten una información totalmente distinta.

Los dos gráficos pertenecen a dos publicaciones diferentes sobre una misma noticia: sostenibilidad del bitcoin mostrando su consumo energético en 2021. En la primera gráfica el bitcoin aparece visualmente situado en el medio después de haber simplificado los datos omitiendo países tanto de mayor o menos consumo que el bitcoin, pero de forma intencionada se han omitido países que estarían por encima. En la gráfica de la derecha aun mostrando el mismo dato sobre consumo de energía del bitcoin, aparecen todos los países que están por encima en consumo y solo uno por debajo, lo que visualmente da la sensación de que el bitcoin está en el puesto más bajo.



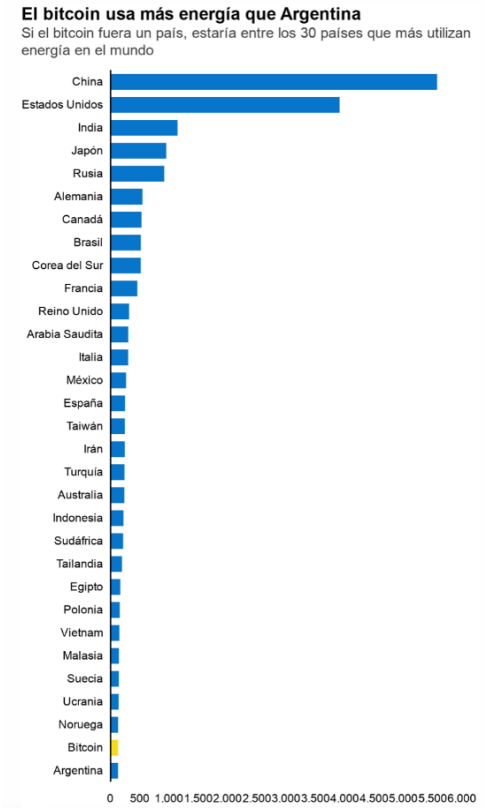


IMAGEN 8 GRÁFICA WEB STATISTA 06/05/2021 **IMAGEN 9 GRÁFICA PERIÓDICO BBC 03/02/2021**

Ningún gráfico puede capturar la realidad en toda su riqueza. Sin embargo, un gráfico puede empeorarla o mejorarla según su capacidad para lograr un equilibrio entre simplificar la realidad y oscurecerla con demasiados detalles.

OCULTACIÓN O CONFUSIÓN POR INCERTIDUMBRE:

Para evitar confusiones los gráficos deben ser precisos, pero en muchos casos demasiada precisión puede ser un obstáculo para la comprensión. Los datos a menudo son inciertos y esa incertidumbre debe ser revelada. Ignorarlo puede conducir a un razonamiento erróneo.

Esta situación se da mucho en visualizaciones sobre datos científicos. La incertidumbre confunde a muchas personas porque se tiene la expectativa de que la ciencia y la estadística descubran verdades precisas, aunque se sabe que lo que hacen es estimaciones imperfectas que pueden estar sujetas a cambios y actualizaciones. Que las actualizaciones sean inciertas no implica que sean erróneas. La palabra error no significa lo mismo que equivocación.

Simplificándolo, un error o incertidumbre significa que cualquier estimación que hacemos sin importar lo precisa que parece en el gráfico es generalmente un valor dentro de un rango de posibles valores. En muchos casos las investigaciones científicas se hacen sobre un cierto rango de incertidumbre que cuando aparece se suele menospreciar. En realidad, mostrar esa incertidumbre es más fiable que obviarla.

En la gráfica que aparece a continuación se observa cómo se ha sombreado la incertidumbre que rodea la estimación. Hubiese sido un fallo muy común no mostrar esa área sombreada.

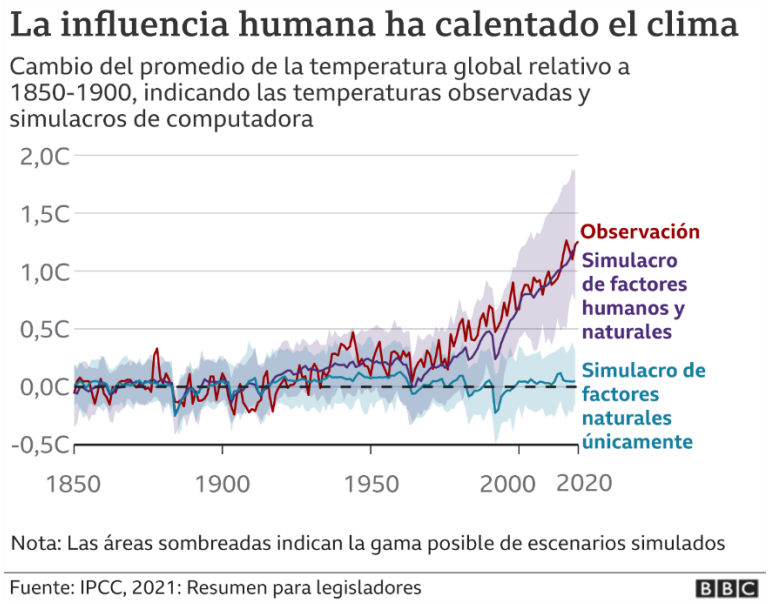


IMAGEN 10 GRÁFICA PERIÓDICO BBC 09/08/2021

SUGERENCIA DE PATRONES ENGAÑOSOS

Como ya se ha mencionado anteriormente, los buenos gráficos son útiles porque simplifican la complejidad de los números. Sin embargo, los gráficos también pueden llevarnos a detectar patrones y tendencias que son ambiguos, principalmente porque el cerebro humano tiene tendencia a interpretar demasiado lo que vemos y tratar siempre de confirmar lo que ya creemos.

Las siguientes tres gráficas sobre el dato de la evolución del el PIB[5] aparecieron el mismo día en tres medios de comunicación. Eligiendo tipos de gráficas diferentes y tomando cada una un periodo de tiempo distinto se obtienen efectos dispares cuando observamos las gráficas.

IMAGEN 11 TELEDIARIO TVE 28/01/2022

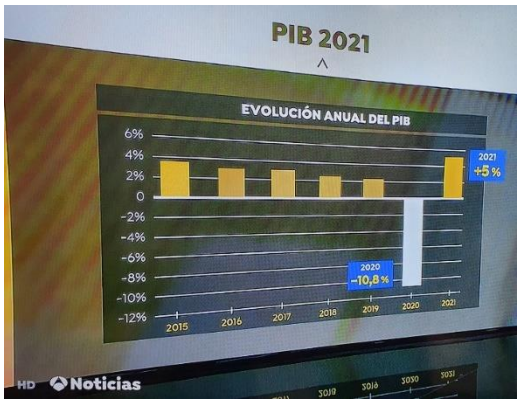


IMAGEN 12 TELEDIARIO ANTENA3 28/01/2022

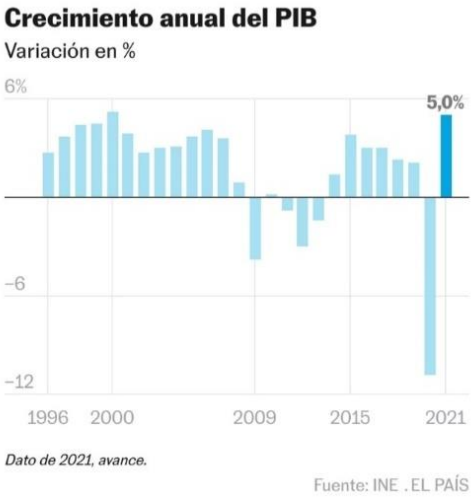


IMAGEN 13 PERIÓDICO EL PAÍS 28/01/2022

[5] PIB: Producto Interior Bruto

Estudio técnico

Contexto tecnológico

La proliferación de las fake news ha hecho que sean muchas las empresas que han puesto en marcha proyectos para proporcionar información de calidad. En 2015 se creó la "International Fact-Checking Network" (IFCN) con el objetivo de reunir a la creciente comunidad de verificadores de hechos (*"fact checkers")* y de defensores de la información de calidad en la lucha global contra la desinformación. La IFCN promueve la excelencia de la verificación de datos en más de 100 organizaciones en todo el mundo a través de actividades de promoción, capacitación y eventos globales. Su principal objetivo es el de controlar las tendencias en el campo de la verificación de hechos para ofrecer recursos, contribuir al discurso público y brindar apoyo para nuevos proyectos e iniciativas. Esta batalla que se está lidiando para dar un valor añadido a las informaciones se ha intensificado durante la pandemia con la que llegó la mayor ola de desinformación que se conoce. El COVID19 creó el entorno perfecto en el que han proliferado la confusión, el desorden informativo y la cantidad excesiva de información que no ayuda a que las personas más vulnerables y viviendo una época tan convulsa, puedan distinguir fuentes confiables y datos fidedignos. Esta nueva situación que nos ha tocado vivir y que es conocida por todos como "infodemia" ha agravado la expansión de noticias y datos inexactos o erróneos y que todos diariamente contemplamos. Por consiguiente, cada vez son más las herramientas que se necesitan para frenar este fenómeno.

En las diferentes áreas se han ido creando con el paso del tiempo y la necesidad de las circunstancias, herramientas específicas para realizar "fact checking" pero cuando hablamos de visualización de datos las herramientas no son lo suficientemente especializadas y las que podrían ayudarnos y ser efectivas son demasiado complicadas.

Algunas de las plataformas que más temas abarcan y con mayor trayectoria están en Estados Unidos: *Snopes.com* y *FactCheking.org*. Pero estas iniciativas se han expandido por otros países como *Factual* en Francia, *Chequeado* en Argentina y *Newtral* en España.

Además, existen webs que nos proporcionan herramientas con una función más específica sobre los diferentes temas que abarca esta disciplina.

Especialmente útiles en el mundo de las redes sociales encontramos:

- Botometer: Detecta usuarios bots en múltiples plataformas.
- twXplorer: Busca en múltiples idiomas todo lo que está sucediendo alrededor de una etiqueta, ofreciéndonos una radiografía sobre ella.
- Twitonomy: Rastrea la operatividad de un perfil de Twitter y da un desglose sobre las publicaciones, seguidores e interacciones de esa cuenta.

Para el análisis de videos y fotos:

- TinyEye: Buscador inverso de fotografías. Rastrea el origen de una imagen, buscando todas las webs donde haya sido publicada.
- InVID: Permite verificar videos y comprobar su contexto, la veracidad del contenido y el cumplimiento de la legislación de derechos de autor.
- Jeffrey's Image Metadata Viewer: Aplicación web que permite conocer los datos de archivo e imagen, como la geolocalización y dispositivo de una fotografía. Podemos descubrir así el origen de la imagen e incluso el tiempo en el que fue tomada.

Navegación en internet:

- Whois: Rastrea los datos de registro de nombre de un dominio, de esta manera facilita la revisión sobre la veracidad de un sitio web, así como la duración en el tiempo y actualizaciones de la página.
- Fact Check Explorer: Plataforma creada por Google que permite hacer una búsqueda rápida sobre bulos publicados alrededor de cualquier persona, entidad, tema, etcétera. La herramienta funciona en varios idiomas (incluido español) y hace una búsqueda rápida en las webs de verificación para saber si alguien ya ha investigado el bulo que nos interesa.

Visualizaciones de datos: Las siguientes herramientas no tienen el objetivo de realizar fact checking pero nos pueden ayudar a extraer información de gráficos que cumplan con unas determinadas características.

- DigitizeIt: Obtiene datos de gráficos de líneas y dispersión. Para ello hay que indicar, sobre la imagen importada, la localización y valores del sistema de coordenadas.
- Macro de Excel: Microsoft proporciona una herramienta para recuperar datos de un gráfico cuando los datos están en una hoja de cálculo o libro externo. Esta funcionalidad es costosa de implementar, pero muy útil en situaciones en las que el gráfico se creó a partir de otro archivo que no está disponible o que se encuentra dañado.

Objetivo

A la vista de que en la actualidad no existe una herramienta que nos ayude en la tarea de realizar el fact checking sobre visualizaciones de datos, el objetivo de este desarrollo será realizar un estudio de viabilidad para entender cuáles podrían ser los pasos para encontrar soluciones que nos ayuden en el proceso de verificación. Para ello se utilizará un enfoque de ciencia de datos. Cada apartado que vaya realizando será un paso en el proceso de comprensión si la gráfica se corresponde con lo que dicen los datos.

Organización

Tras el estudio pormenorizado de la situación actual del fact checking, de la relevancia de las gráficas y de cómo mediante el uso de las mismas nos podemos desinformar he creído conveniente enfocarlo como describo a continuación: en primer lugar, es necesario realizar un modelo de clasificación de gráficas previo a cualquier otro tratamiento puesto que la forma de interpretar las gráficas depende del tipo de las mismas. Por ello el primer paso será aplicar técnicas de visualización por computador para clasificar las gráficas. A continuación, se buscará aplicar técnicas de procesamiento de imágenes que nos ayuden en el proceso inverso al que se hace habitualmente al crear una gráfica. Normalmente a partir de datos se construye una gráfica, ahora nuestro problema es a partir de una gráfica extraer sus datos.

Para abordar este estudio las tecnologías de las que se hará uso en el proyecto son las siguientes:

- Python es el idioma sobre el que realizaremos el análisis
- Cuadernos de Jupyter como medio de desarrollo haciendo uso de la plataforma Google Colab
- Librería Google images download para la descarga de imágenes
- Github como repositorio para el código
- Librerías OpenCV, Matplotlib, Scikit-learn, PyTorch, SKIMAGE

Clasificación de gráficas

Las gráficas no tienen un lenguaje común; se componen de marcadores, escalas y leyendas que interpretaremos de una forma diferente según el tipo de gráfico que los contenga. Por lo tanto, el primer paso de este desarrollo será implementar un modelo de clasificación mediante aprendizaje supervisado.

Obtención y tratamiento del dataset

Para la realización de un modelo de clasificación supervisada es necesario tener un conjunto de datos etiquetados. Para la obtención del dataset[6] he utilizado la librería "Google images download" que permite la descarga de imágenes resultantes de una búsqueda de Google. La utilización de esta librería me dio la opción trabajar con un conjunto de datos donde las imágenes están etiquetadas según el directorio que las contiene. Además, en un futuro si el desarrollo crece, permitiría obtener de forma sencilla y rápida nuevas categorías de imágenes.

Debido a la amplia variedad de tipos de gráficas he decidido centrarme en aquellos que durante la fase de análisis del proyecto me he encontrado con mayor frecuencia: gráficos de barras, sectores, líneas y dispersión.

[6] Anglicismo comúnmente utilizado en la informática para referirse a un conjunto de datos.

```
from google_images_download import google_images_download

response = google_images_download.googleimagesdownload()
for tipoGrafica in ["Bar chart", "Pie chart", "Line graphs", "Scatter plots"]:
  arguments = {"keywords": tipoGrafica,"limit":100, "format":"png","print_urls":True}
  paths = response.download(arguments)
  print(paths)
```

IMAGEN 14 CÓDIGO PARA LA OBTENCIÓN DEL DATASET

Modelo de clasificación

Una vez obtenido el dataset lo he dividido en dos: un 70% se va a utilizar para el entrenamiento del modelo y el 30% restante se utilizará para el testeo.

Como el dataset proviene de una búsqueda de Google las imágenes que incluye son muy variadas y cada una con un tamaño diferente. El DataBlock[7] que usaré para consumir los datos para entrenar mi modelo también me permitirá aplicarle un preprocesado a las imágenes: el escalado para que tengan un tamaño homogéneo y la introducción de ruido gaussiano. Con el uso de la técnica de ruido gaussiano busco degradar sutilmente la calidad de las imágenes, siendo en muchas ocasiones prácticamente imperceptible para el ojo humano pero para los ordenadores supone estar frente a una imagen diferente. Consigo así un modelo más robusto y menos susceptible a la calidad de las imágenes de entrada. Esta transformación no se encontraba dentro de las opciones que por defecto ofrece la librería por lo que tuve que definirla como vemos en la imagen 16.

```
db = DataBlock(blocks = (ImageBlock, CategoryBlock),
                get_items=get_image_files,
                splitter=RandomSplitter(valid_pct=0.2,seed=42),
                get_y=parent_label,
                item_tfms =[ Resize(256), AddNoise(mean=0., std=100.)],
                batch_tfms=aug_transforms(size=128,min_scale=0.75))
```

IMAGEN 15 TRATAMIENTO DE LAS IMÁGENES DE ENTRENAMIENTO

```
class AddNoise(RandTransform):

    def __init__(self, mean=0., std=1., **kwargs):
        self.std = std
        self.mean = mean
        super().__init__(**kwargs)

    def encodes(self, x:TensorImage):
        return x + torch.randn(x.size()) * self.std + self.mean
```

IMAGEN 16 FUNCIÓN PARA AÑADIR RUIDO A UNA IMAGEN

[7] DataBlock: Objeto genérico para poder consumir de forma estructurada Datasets y que permite aplicar transformaciones a los datos. https://pytorch.org/tutorials/beginner/basics/data_tutorial.html

Para entrenar el modelo elijo una red neuronal convolucional pre-entrenada (CNN) con el modelo para clasificación de imágenes *modelResnet18*. Me decanto por este modelo por su eficiencia y precisión (como se puede ver en la *imagen 18*). Aplico la técnica de *fine tunning* que consiste en desarrollar un modelo concreto para mi problema a partir de un modelo pre-entrenado. Estos modelos han sido previamente entrenados con datasets muy grandes que requieren una alta capacidad computacional y de tiempo. A este modelo le añado las capas necesarias para adaptarlo a mi problema y así mejorar notablemente los resultados.

```
learn = cnn_learner(dls,resnet18,metrics=accuracy,cbs=callbacks).to_fp16()
```

IMAGEN 17 DECLARACIÓN DEL MODELO DE CLASIFICACIÓN

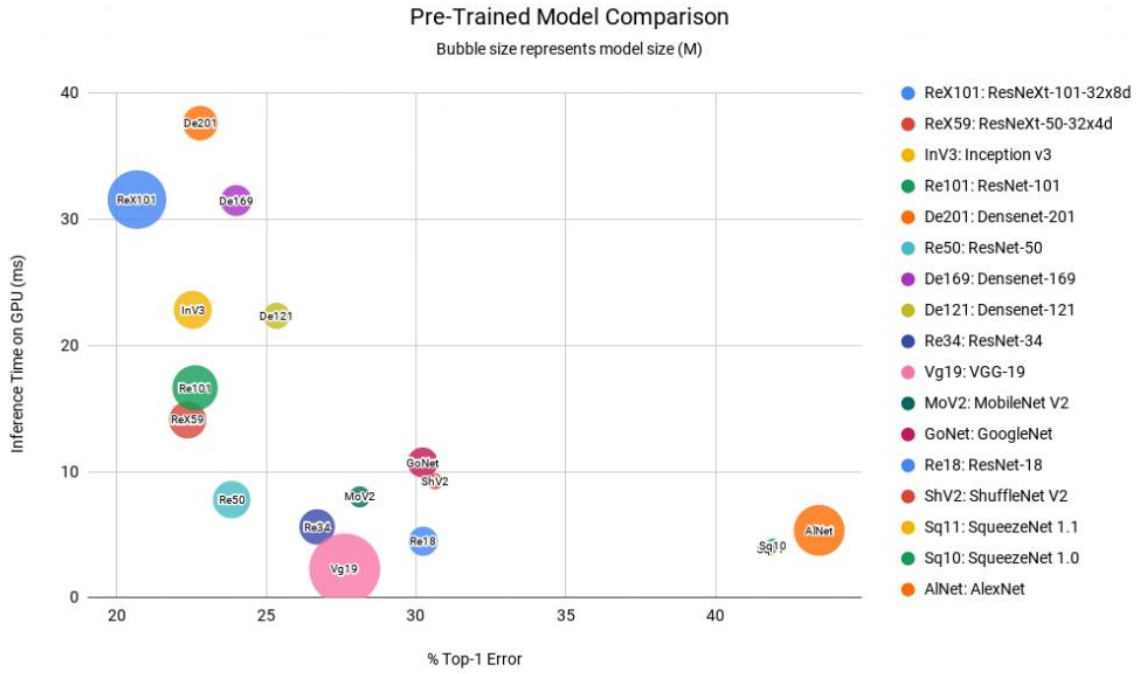


IMAGEN 18 COMPARATIVA DE MODELOS PRE-ENTRENADOS

Con esta configuración el modelo obtiene resultados muy satisfactorios con una "accuracy" (precisión) de 0.*955224* y la siguiente matriz de confusión.

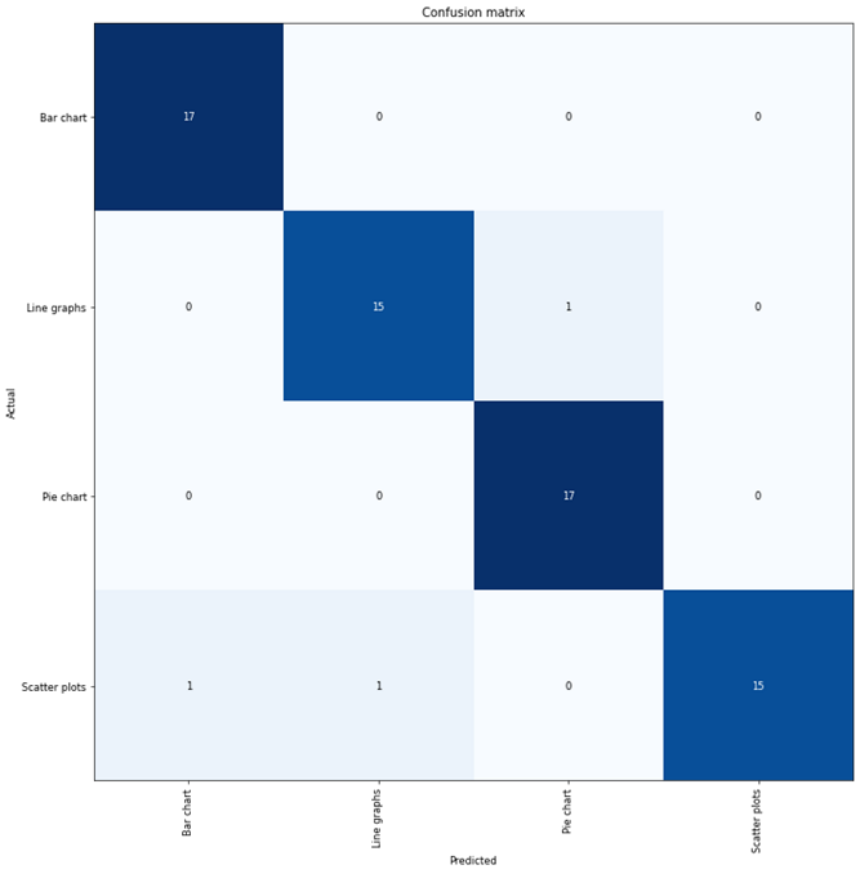


IMAGEN 19 MATRIZ DE CONFUSIÓN DEL MODELO DE CLASIFICACIÓN

De esta forma he conseguido que el modelo sea capaz de clasificar una imagen dentro de los cuatro tipos de gráficas con las que se trabaja.

Método de obtención de datos

Una vez que he podido clasificar una gráfica, el siguiente paso es extraer los datos numéricos que nos muestra. Cada tipo de gráfica clasificada anteriormente necesita procesar la información visual de forma diferente. Después de valorar las categorías de gráficas con las que he decidido trabajar y el alcance al que puede llegar este proyecto, he decidido extraer la información aportada en una visualización de datos a través de marcadores de áreas (en este caso los gráficos de sectores).

Para este proceso valoré en un primer momento el uso de segmentación semántica o segmentación por instancias, pero teniendo en cuenta factores como el tiempo y la complejidad de estos modelos opté por técnicas más sencillas que desarrollo a continuación.

En un principio, las pruebas que realicé fueron utilizando el algoritmo de agrupamiento de datos DBSCAN[8] que es un algoritmo de clustering no supervisado basado en la densidad. Al ejecutarlo con un cuaderno de *Jupyter* a través de la herramienta *Google Colab* los tiempos de cómputo eran demasiado largos. También probé con el algoritmo

[8] Librería Scikit learn https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

OPTICS[9] parecido a DBSCAN pero con un consumo de RAM mucho menor. Al ver que el problema del tiempo de ejecución persistía decidí cambiar la estrategia.

A partir de este punto decidí documentarme sobre la librería de Python "matplotlib" y probar qué herramientas me proporcionaba. Matplotlib es una librería para crear visualizaciones estáticas e interactivas en Python.

Hallé que la información que necesitaba podía obtenerla mediante los contornos de las áreas que representaba cada dato. Si obtenía la proporción que ocupaba cada sector de la circunferencia tendría los porcentajes que estaban representando. Sin embargo, como se puede apreciar en la siguiente imagen este método no era lo suficientemente preciso.

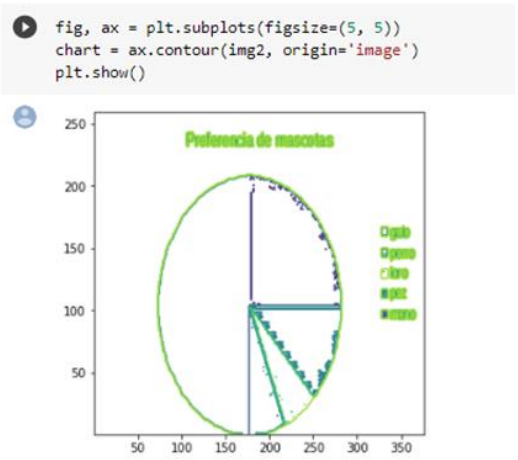


IMAGEN 20 CONTORNOS DETECTADOS EN LA GRÁFICA

Como se puede observar en la imagen no todos los contornos se detectaban de forma correcta. Para resaltar los contornos antes de intentar obtener las áreas que comprende cada uno probé las siguientes técnicas de procesamiento de imágenes:

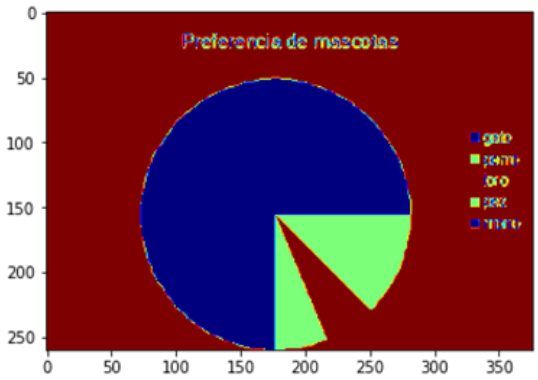


IMAGEN 21 TÉCNICA "THRESHOLD_MULTIOTSU"

[9] Librería Scikit learn https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html?highlight=optics#sklearn.cluster.OPTICS

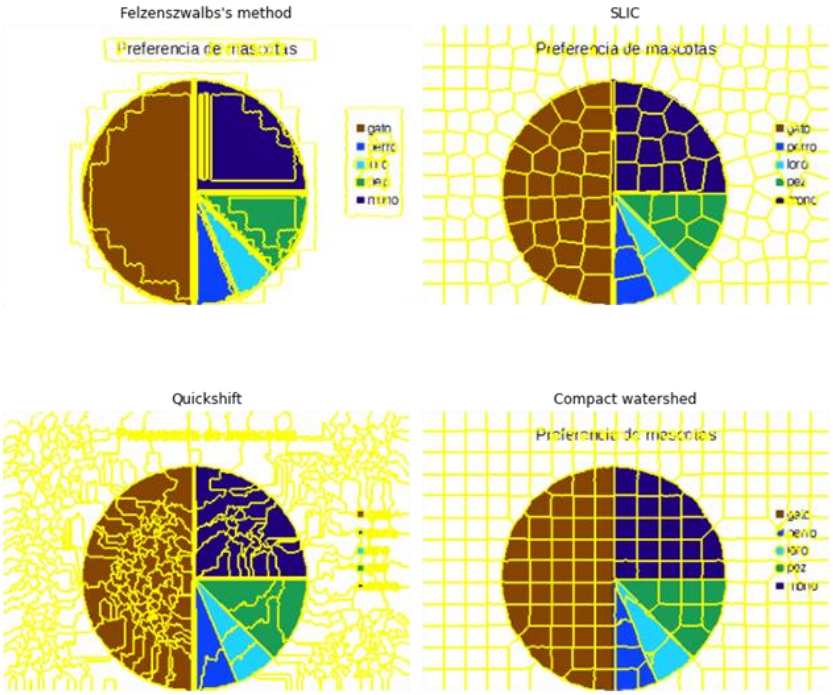


IMAGEN 22 LIBRERÍA "SKIMAGE.SEGMENTATION"

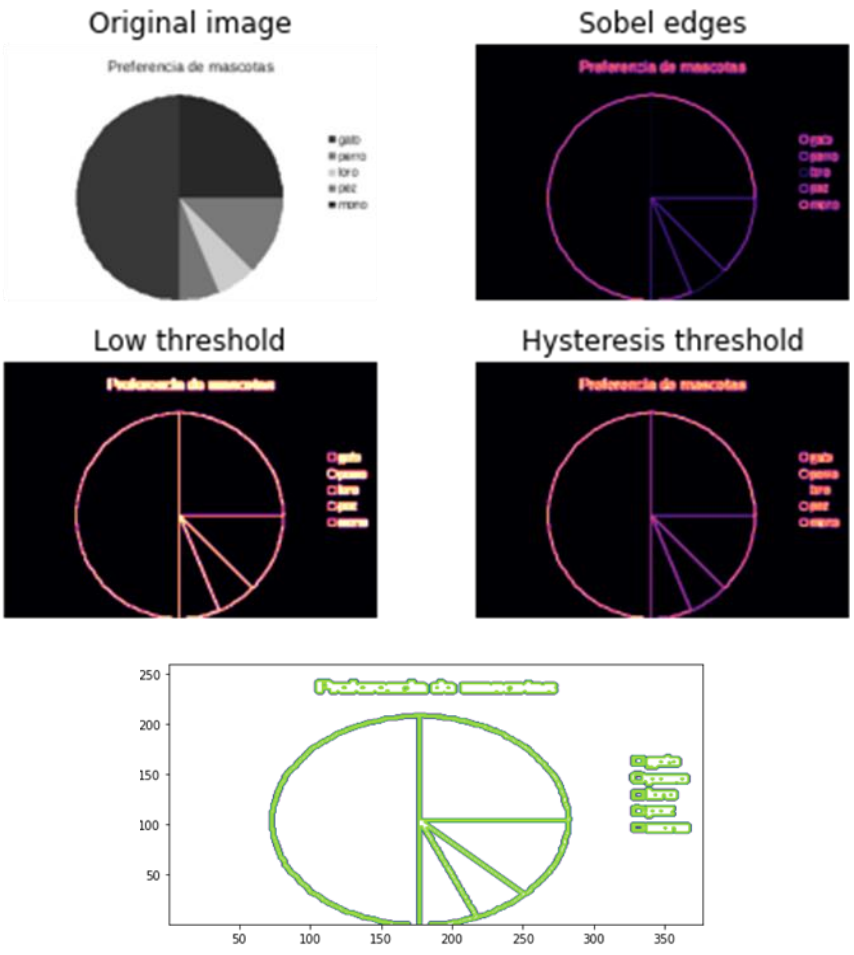


IMAGEN 23 LIBRERÍA "SKIMAGE FILTERS"

Aplicando estas técnicas lo que pretendía era facilitarle a la librería la detección de los bordes. Con los bordes más definidos conseguí tener cualquier sector de cualquier gráfica perfectamente delimitado. De esta forma, pude dividir la imagen en áreas que se encuentran en capas, y por el polígono que forman pude obtener las áreas que necesitaba (como se aprecia en la imagen inferior). Me encontré con el problema de que estas capas no solo contenían los datos que me interesaban, sino también textos e información desordenada. Puesto que esta información no se estructura igual para todas las gráficas la selección de contornos tampoco. Logré los datos de las proporciones que representa un gráfico de pruebas, pero el proceso de automatizar la selección de las áreas correctas sería demasiado costoso.

```
from shapely.geometry import MultiPolygon
contours = map(np.squeeze, chart.allsegs[1])
polygons = map(Polygon, chart.allsegs[1])
multipolygon = MultiPolygon(polygons)
multipolygon
```

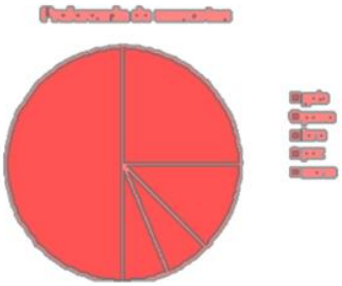


IMAGEN 24 CONTORNOS OBTENIDOS

Intenté esta misma técnica de obtener las áreas con algoritmos de la librería OpenCv. Como se puede apreciar en las siguientes imágenes, al aplicar a la imagen inicial técnicas para definir los contornos los resultados mejoraron notablemente, pero al igual que me ocurría anteriormente no pude seleccionar de forma automática las áreas que me interesaban.

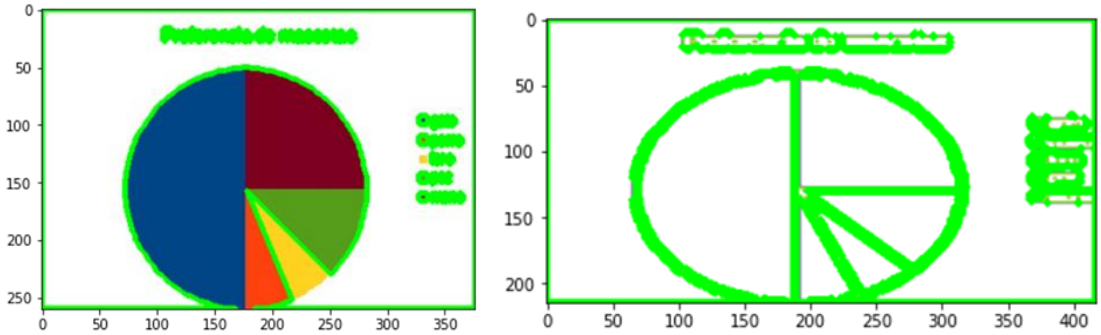


IMAGEN 25 DETECCIÓN DE CONTORNO DE LA LIBRERÍA OPENCV

Tras este proceso cambié la técnica intentando obtener resultados más fáciles de procesar. Recurrí nuevamente a las funciones de la librería *matplotlib* para unificar los colores de cada sector, después contar la cantidad de pixeles de cada color y finalmente obtener las proporciones.

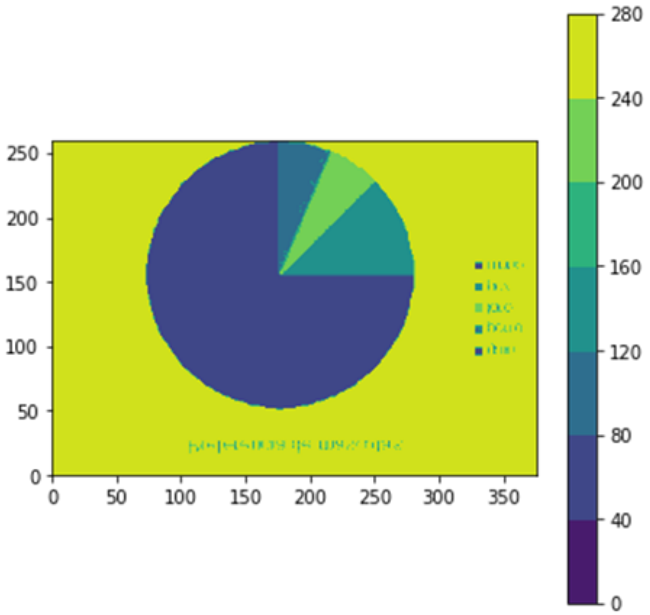


IMAGEN 26 PROPIEDAD COLORBAR DE MATPLOTLIB

Conseguí obtener datos de la cantidad de pixeles por colores pero el problema de la automatización persistía.

Continuación del desarrollo

El alcance del proyecto no ha conseguido automatizar el proceso de obtención de datos genérico para cualquier gráfica de sectores. Por consiguiente, el siguiente paso sería conseguir crear esa función que he estado buscando y que sirva para cualquier gráfica de sectores. Así como investigar otras técnicas, por ejemplo, la segmentación semántica que descarté en un primer lugar.

La misma técnica se podría aplicar a otro tipo de gráficos que representen la información a través de áreas, como por ejemplo el gráfico de rectángulos. El método sería válido para los gráficos de barras si lo adaptamos para que obtenga el valor de la barra de mayor tamaño y calcule la proporción para el resto.

Un paso más en el desarrollo sería relacionar los diferentes métodos para cada tipo de gráfica con el modelo de clasificación inicial. De esta forma le pasaríamos una gráfica de cualquier tipo a nuestro modelo y él la clasificaría e interpretaría de forma correcta.

Esta podría ser la base de un modelo que puede crecer tanto como clasificaciones de gráficas se añadan.

Conclusiones

La realización de este trabajo fin de máster ha sido un desafío tanto a nivel personal como en la faceta técnica. En este proyecto he analizado la importancia que tienen hoy en día las gráficas como medio de difusión de la información, por un lado trasmiten información de forma visual y por otro lado nos permiten tratar con grandes cantidades de datos. Tras el análisis me he percatado de la relevancia que tiene saber interpretar una gráfica correctamente porque al mismo tiempo que informan pueden convertirse en herramientas de desinformación.

He partido de un marco teórico que aborda el concepto del fact checking, la importancia del mismo en la actualidad, y la relevancia que tiene o que debería tener en la visualización de datos como método que se nos presenta eficaz para entender apropiadamente el volumen de datos que visualizamos a diario.

La tecnología en algunos ámbitos de la información nos proporciona herramientas especializadas para ayudarnos en la verificación de los contenidos ante una realidad compleja. Pero en el caso de las visualizaciones de datos la inexistencia de investigaciones previas y la dificultad añadida que supone automatizar la interpretación de las imágenes ha propiciado la existencia de un vacío en este campo que me ha supuesto un gran reto en el planteamiento del desarrollo técnico del proyecto.

A nivel personal este trabajo ha sido un reto por las características del mismo. En mi día a día tanto en la etapa de estudios como en la profesional estoy acostumbrada a trabajar con un objetivo y todo mi trabajo gira entorno a la consecución de dicho objetivo. Debo analizar, planificar y realizar mis tareas siempre con la vista puesta en el fin propuesto. En este trabajo el planteamiento ha sido diferente porque mi punto de partida era un tema determinado sin un objetivo claro y del que no tenía mucha noción. Mi labor ha consistido en tratar de informarme, leer, investigar, y buscar la manera de enfocar el tema para poder aplicar los conocimientos adquiridos.

Bibliografía

Libro "How chats lie" de Alberto Cairo.

Libro "Gramática de las gráficas" de Joaquín Sevilla Moróder.

Información sobre el uso de datos: https://www.domo.com/learn/infographic/data-never-sleeps-8

Gráficas de datos de La Rioja https://bi.larioja.org/pentaho/covid19

http://www.aikaeducacion.com/recursos/fact-checking-para-combatir-la-desinformacion/

https://www.tupreicfesinteractivo.com/2011/07/como-leer-un-grafico.html

Documentación de modelos pre-entrenados https://learnopencv.com/pytorch-for-beginners-image-classification-using-pre-trained-models/

Documentación de la librería fastai https://docs.fast.ai/

Documentación de la librería para descargar imágenes de Google https://google-images-download.readthedocs.io/en/latest/index.html

Documentación de la librería PyTorch https://pytorch.org/

Documentación para la obtención de la función de ruido del datablock https://forums.fast.ai/t/tutorial-on-adding-new-data-augs-gaussian-noise/81198

Documentación de la librería OpenCV https://docs.opencv.org/