Dataton 2022

Actualmente en el auge de las nuevas tecnologías y la industria 4.0, podemos observar como necesidades que generalmente son tediosas y repetitivas pueden ser suplidas por el uso de estas tecnologías, en el marco del Dataton 2022 realizado por Bancolombia se plantea el reto de generar un recomendador de noticias para sus gerentes de campo, los cuales deben interactuar continuamente con sus clientes, esto con el fin de agilizar y amenizar las conversaciones con dicho cliente, generar contextos rápidos de la situación de su cliente o del sector sobre el que se mueve para propiciar interacciones enfocadas a la empatía y bienestar de este.
Para poder abordar dicho problema el equipo Goofy Goobers ha aceptado el reto de la Dataton, primero se inicia realizando un preprocesamiento de los datos (noticias) ya que los datos no están etiquetados, por otra parte tampoco han pasado un filtro que garantice o permita identificar si son noticias “Fake” o no, de forma que se plantea el uso inicial de dos modelos para abordar estos segmentos. Utilizando un pipeline de hugging face una comunidad de inteligencia artificial que reúne los aportes más representativos de ingenieros de diferentes disciplinas para la solución de problemas de diferentes naturalezas, entre ellos:

- Zero Shot Pipeline: Línea de clasificación basada en NLI y Zero-Shot utiliza un ModelForSequenceClassification entrenado en tareas de NLI (inferencia del lenguaje natural). Se puede pasar cualquier combinación de secuencias y etiquetas y cada combinación se planteará como un par de premisas/hipótesis y se pasará al modelo pre entrenado. El modelo pre entrenado seleccionado para trabajar el proceso de clasificación se llama bert-base-spanish-wwm-cased-xnli
- Bert Base Spanish es una versión de BERT el modelo de NLP que utiliza la arquitectura de Transformers para leer, clasificar, sugerir o generar texto de forma dinámica, aprendiendo de forma optimizada y ofreciendo muy buenos resultados, originalmente, BERT empezó únicamente para el inglés pero se ha extendido a otros idiomas como el español con su versión BETO que ha sido entrando con el mayor corpus de español conocido hasta la fecha.
- fake-news-detection-spanish es un modelo de detección de noticias falsas en español, esta permite identificar si una noticia es falsa o real, se crea a partir de roBERTa. RoBERTa-large-bne es un modelo de lenguaje enmascarado basado en transformadores para el idioma español. Está basado en el modelo RoBERTa large y ha sido preentrenado utilizando el mayor corpus de español conocido hasta la fecha, con un total de 570 GB de texto limpio y deduplicado procesado para este trabajo, recopilado a partir de los rastreos web realizados por la Biblioteca Nacional de España entre 2009 y 2019.

Metodología:

El proyecto usa como base tres archivos CSV con los nombres clientes.csv, noticias.csv y clientes-noticias.csv donde encontramos información relacionada a los clientes del banco, un conjunto de noticias (23377) obtenidas de diferentes fuentes y una relación entre noticias y clientes, por medio de estos documentos empezamos las siguientes etapas de trabajo:

Preprocesamiento:

Al observar los documentos y como se platicó en la apertura del evento, los datos en noticias.csv son datos no clasificados, cuentan con un filtro de limpieza realizado para poder trabajar de forma uniforme con ellos, pero no cuentan con las etiquetas que se piden para la estructuración del recomendador, estas etiquetas son:

- Macroeconomía: con un nivel de importancia 2.
- Sostenibilidad: con un nivel de importancia 1.8.
- Innovación: con un nivel de importancia 1.7.
- Regulaciones: con un nivel de importancia 1.6.
- Alianzas: con un nivel de importancia 1.4.
- Reputación: con un nivel de importancia 1.2.
- Otra: con un nivel de importancia 0.5.
- Descartable: con un nivel de importancia 0.

Sin embargo también se ha identificado la necesidad de verificar si las noticias pueden o no ser fake news (noticias falsas) de forma que antes de iniciar la clasificación se realiza el filtro de fake news.

- Preparación de los datos: Para utilizar el filtro de Fake news se utiliza un conjunto de ciclos anidados que van a descomponen las noticias en pasos o fragmentos de 500 caracteres de forma que puedan utilizarse en el modelo, ya que el modelo de fake news tiene un límite de procesamiento de 500 caracteres, de forma que el ciclo genera en últimas un df dividió por fragmentos de la noticia con su respectivo id para permitir identificar este

index	id	fragmento
0	news10006	Precio dolar hoy: la cotizacion de la divisa aumento y supero los $4.300. [SEP] Este martes, 2 de agosto, el dolar alcanzo un precio maximo de $4.328 y un minimo de $4.245. Ademas, se negociaron US$889,35 millones en 1.650 transacciones. Un cambio en la divisa que representa una subia de $68,55 frente a la Tasa Representativa del Mercado (TRM hoy), que es de $4.314,54. TRM hoy: siga la cotizacion del dolar, en vivo, hoy viernes 22 de julio Le puede interesar: Cinco recomendaciones para hacerle e
1	news10006	l quite al precio del dolar y a la inflacion Escuche el podcast El Mundo Hoy
0	news10011	Es cierto: El presidente Gustavo Petro quiere que se acabe el detestable impuesto del 4 x 1000. Aqui le contamos como seria [SEP] No hay nada mas permanente que lo temporal y asi ocurrio con el impuesto del 4 x 1.000 con el que se gravaron algunas operaciones financieras. Aunque multiples veces se han radicado proyectos de ley en el Congreso de la Republica para desmontarlo, hasta el momento ninguno ha tenido exito y ya tiene 24 anos de vigencia en Colombia. Este impuesto, tambien conocido como
1	news10011	gravamen a los movimientos financieros (GMF), nacio para hacerle frente a la crisis bancaria por la que estaba pasando el pais en 1998 y, en ese entonces, se esperaba que su vigencia fuera solo por un ano. Sin embargo, su eliminacion se fue postergando con el pasar de los anos porque se convirtio en una fuente de financiacion para la reconstruccion de zonas afectadas por el terremoto del Eje Cafetero (1999) y ayudo a enfrentar la crisis de 2014 por el desplome de los precios de petroleo, que afe
2	news10011	cto fuertemente los ingresos de la nacion. Al parecer, despues de 24 anos de vigencia este impuesto podria llegar a su fin en Colombia , pero no para todos los colombianos, de acuerdo con una propuesta que fue incluida en el proyecto de ley de la reforma tributaria que radico el Ministerio de Hacienda el pasado lunes 8 de agosto . De acuerdo con el director de Impuestos y Aduanas Nacionales (Dian), Luis Carlos Reyes, si se aprueba el articulo 64 de la reforma tributaria, se acaba el 4 x 1.000 pa
3	news10011	ra quienes hagan movimientos menores a $13,3 millones al mes. "Ya existe una exencion para una sola cuenta bancaria marcada. En adelante, los bancos deberan consolidar informacion entre ellos para que la exencion aplique sin necesidad de marcar la cuenta", dijo. Cuando se creo, este impuesto era de solo un 2 x 1.000 , es decir, que se debian pagar $2 por cada $1.000 de alguna operacion financiera. Luego se incremento a un 3 x 1.000 y en 2006 subio al 4 x 1.000 actual, que se cobra, por ejemplo,
4	news10011	por los retiros en cajeros o ventanilla, desembolsos de creditos o transacciones en cuentas corrientes o de ahorro . Contrario a lo que muchos colombianos piensan, aunque son las entidades financieras las encargadas de recaudar este impuesto del 4 x 1.000, este dinero finalmente llega a las arcas de la nacion para financiar el Presupuesto General de cada ano. Es decir, que no es un dinero con el que se queden los bancos. De acuerdo con datos de la Dian, con corte a abril de 2022,

- Ejecución del modelo fake news: El modelo fake news es un modelo descargado por la pipeline de zero-shot classification de hugging face community este permite generar un label junto a un score que determina si una noticia es más o menos problema que sea real o fake. Los datos de entrada del modelo deben componerse por

Título_de_la_noticia + [“SEP”] +cuerpo_de_la_noticia

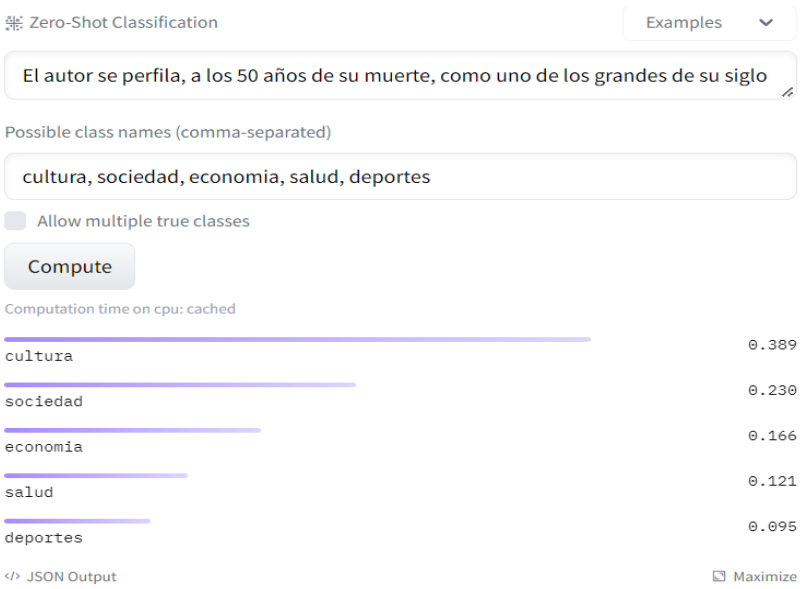
y este arroja como salida un diccionario con un conjunto llamado “labels” y otro llamado “score” estos son los que nos permiten conocer la probabilidad de que una noticia sea “REAL” o “FAKE”, de forma que se pasa el data frame con los fragmentos por el modelo.

	label	score
0	REAL	0.999487
1	FAKE	0.999920
2	FAKE	0.999975
3	FAKE	0.999973
4	FAKE	0.999969
...	...	...
995	FAKE	0.999488
996	FAKE	0.999537
997	FAKE	0.999969
998	FAKE	0.999765
999	FAKE	0.991642

Por último se observa el promedio de los resultados y los porcentajes según el porcentaje, en nuestro caso decidimos que el threshold estaría marcado para aquellos que tuvieran un porcentaje mayor a 75 por ciento para la etiqueta “FAKE” y se realiza una promediación por etiquetas con el objetivo de realizar una doble validación para asegurar el etiquetado de forma que las noticias con la etiqueta “fake” serán descartadas directamente con la etiqueta descartables.

- Modelo de clasificación: EL modelo de clasificación utilizado es una variación de BETO la versión española de BERT, el cual ha sido entrenado con un corpus de texto de más de 500 gb, la versión seleccionada es una variante de BETO tuneada utilizando de base XNLI con un accuracy de 79,9 el cual permite obtener un valor probabiliste de las categorías a las cuales se relaciona más el texto introducido.

- Preprocesamiento: Para utilizar Recognai/bert-base-spanish-wwm-cased-xnli se generó un csv de sinónimos relacionados a las etiquetas, formando str vectores que permitirán agilizar los procesos de clasificación para así reducir la fuga de datos relacionados a estos, con la sorpresa de que el modelo era capaz de recibir no solo palabras sino vectores de palabras, evitando tener que utilizar recursos de repetición o ciclos de acción del modelo.



- Ejecución del modelo Recognai/bert-base-spanish-wwm-cased-xnli: Para correr el modelo debe importarse este desde el pipeline, además de solicitar con el hiper parámetro device=0 el modelo que usa GPU para agilizar procesos, posteriormente la estructura que entra al modelo es la siguiente:

  Classifier(“texto”, “vectores de palabras o palabras”)

  De esta el modelo arroja un diccionario con los labels y score de estos mismos donde indica la probabilidad de que esa categoría es a la que pertenece la noticia, posteriormente se crea un df con los valores obtenidos por el modelo y se anexan también los ID para identificar las noticias.

Hasta el momento se ha podido realizar el filtrado tanto en valores de Fake news para garantizar veracidad y responsabilidad, además de usar modelos basados en transformers como son BETO y RoBERTa.