Datatón Bancolombia 14/11/2022

NATA (Noticias de actualidad transparentes y asertivas)

Goofy Goobers Team- Daniel Arias Garzón & Joshua Bernal Salcedo

Introducción

Siendo caracterizada como una empresa en la vanguardia de las tecnologías 4.0, Bancolombia, desarrolla los retos datatón desde 2017, en búsqueda de talento profesional y soluciones a problemas puntuales o situaciones de mejora que enfrentan a nivel interno, en esta versión del datatón 2022 se plantea un reto relacionado a la gestión de noticias, con el objetivo de mejorar la calidad de información que manejan los gerentes de campo, los cuales se enfrentan continuamente a retos y a interacciones con sus clientes donde la información toma un factor clave, de forma que necesitan estar al tanto de todos los cambios o situaciones que se viven en la actualidad del país, teniendo en cuenta lo anterior se plantea la necesidad de crear un sistema inteligente que realice la recomendación de las noticias más relevantes, esto siguiendo unos pesos de importancia

El objetivo es generar un sistema que teniendo el contenido de la noticia, se tenga en cuenta si esta implica directamente al cliente, o indirectamente (a nivel del sector donde se desarrolla la actividad económica), ofreciendo al gerente las noticias más importantes en relación con este, de manera que pueda implementar diferentes ítems que le permitirán al gerente mantener un nivel de calidad excelente en sus servicios.

Por este motivo, el equipo Goofy Goobers asumió el reto creando el sistema NATA, un sistema de noticias diarias que permite entregar al gerente un conjunto de recomendaciones basadas en el cliente y el sector donde se desempeña este.

¿Qué es NATA?

Es una propuesta alternativa que busca implementar tecnología 4.0, haciendo uso de diferentes métodos y estructuras como los transformers, en concreto sistemas basados en BERT (Representaciones de codificadores bidireccionales de transformers).

Técnicamente, BERT es un modelo de Redes Neuronales Artificiales (RNA) aplicado al campo del Natural Language Processing (NLP), específicamente al subcampo del Natural Language Understanding (NLU). BERT en mayor medida, aplican un método de análisis que permite contextualizar de forma más natural cada consulta.

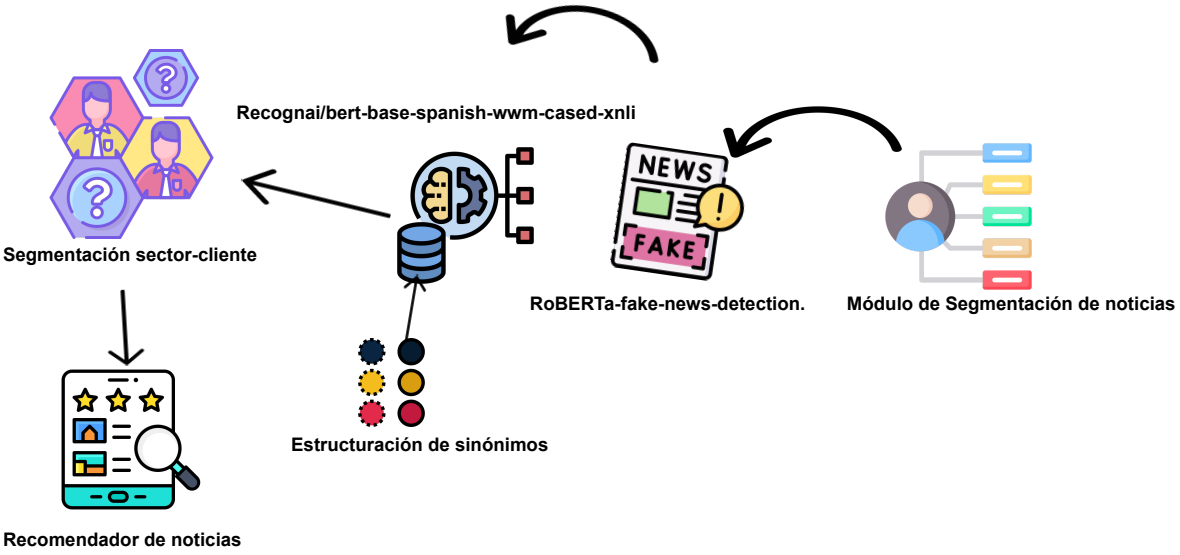
Para lograrlo, BERT posee una característica que se llama “bidireccionalidad”, que consiste en analizar una oración en dos direcciones. Es decir, analiza las palabras que se encuentran tanto a la izquierda como a la derecha de una palabra clave, y esto le permite entender en profundidad el contexto y la temática de toda la frase que introduce un usuario para la búsqueda en Google. Si bien el origen de BERT está especificado para su funcionamiento en textos de habla inglesa, sin embargo, la continua actualización y mejora de partes de este, ha permitido generar variantes como BETO, ELMO y alBERTa, que permiten hacer clasificación y análisis contextual de los textos, de forma que permiten identificar su relación con palabras de tipo “marcador” y generar una clasificación por resultados probabilísticos.

BETO y alBERTa son variantes de BERT que funcionan por medio de un corpus en español, pero no cualquier corpus sino el más grande hasta ahora, recogiendo más de 500 gb de texto en dicho idioma, además de que funciona con base en la librería NLTK que permite tokenizar las palabras de diferentes formas, sin embargo se enfoca en el uso de las funciones Word_tokenize y Punkt_tokenize, las cuales hacen tokenización por palabra o punto a punto.

NATA nace como un sistema estructurado por diferentes modelos que son gestionados mediante un pipeline de zero_one_shot de huggingface una de las comunidades más grandes que trabajan IA y sistemas relacionados a los transformers.

Zero Shot Pipeline: Es la 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, donde 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]

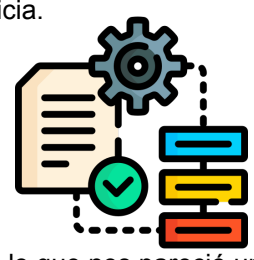
NATA se compone de diferentes modelos y estructuras como:



Metodología

El proyecto tiene una base empírico-analítica que usa como fuente tres archivos CSV con los nombres clientes.csv, noticias.csv y clientes-noticias.csv donde encontramos información relacionada a los clientes del banco y un conjunto de noticias (23377) estas se crearon a partir de un proceso de web scraping en google noticias recogiendo textos de diferentes fuentes de noticia.

Preprocesamiento y tokenización



Para iniciar esta etapa se cargó el documento de noticias.csv, de forma que iniciamos un proceso de inspección de los datos, donde encontramos que existían noticias repetidas aun cuando era, de diferentes fuentes, su estructura era similar por no decir la misma. También encontramos noticias de tendencia, lo que nos pareció un aspecto relevante a tener en cuenta para el proceso de recomendación. Sin embargo en el preprocesamiento pudimos hallar aspectos importantes que afectarían significativamente resultados y los futuros procedimientos, estos fueron:

- El filtrado realizado por los responsables de la creación de la base de datos fue más un aspecto negativo que positivo, el quitar tildes y virgulillas en los textos cambiaban su significado y contexto, generando cambios en palabras como año→ano y el sufijo sión→sion entre otros.

- Fallos en el web scraping: utilizar técnicas de web scraping es complejo, especialmente cuando se desea la mayor pureza relacionada a los datos que se quieren obtener del crudo en los códigos HTML de las páginas de donde se obtiene dicha información, uno de los fallos relacionados a dicha complejidad observada en los datos de *"noticias"* fue que los textos obtenidos en el proceso, no eran netamente noticias, un conjunto grande de estos eran textos donde se orientaban a titulares de noticias obtenidos de páginas de radio, otras eran conjuntos de datos como horas, enlaces, direcciones de ruta o se encontraban tablas que dificultan la aplicación de los modelos, de forma que se trató de detectar la mayor cantidad de noticias y retirarlas con la función .drop, sin embargo esto hizo que la pureza de los datos estuviera sujeta a la duda y de ser rigurosos generaría fallos en procesos de auditoría al tener en cuenta que los datos están sujetos a los sesgos nombrados anteriormente.

- Se realizó un análisis de las palabras tokenizadas con referencia a su peso, encontrando que a la hora de usar las función por defecto de tokenización esta generaba diferentes pesos a palabras bien y mal escritas, lo que incentiva la idea de implementar modelos basados en textos de correcta ortografía facilitando el contexto y la interpretación para el modelo.

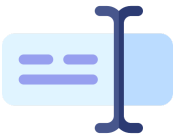
```
tokenizer('maximo')
```

```
{'input_ids': [0, 43889, 2], 'attention_mask': [1, 1, 1]}
```

```
tokenizer('máximo')
```

```
{'input_ids': [0, 3498, 2], 'attention_mask': [1, 1, 1]}
```

Tokenización



Antes de utilizar el modelo para filtrar las fake news en el dataset, se tuvo en cuenta las especificaciones que este establecía para ser usado, entre ellas existe un límite de caracteres (570) el cual requería modificar la composición de los textos de forma que estos por pequeños segmentos coherentes y ordenados pasarán por el filtro. Para ello se buscó realizar una tokenización de los textos con la función PunktSentenceTokenizer de NLTK que permite tokenizar textos de punto a punto, de forma que se recorren los fragmentos tokenizados, además de eso se implementó la función AutoTokenizer la cual codifica las palabras en valores numéricos, esto permitió que las segmentación fuera más agil, y verificar situaciones de fallo y error, ya que se pudo observar como al tokenizar palabras como atención→atencion los valores de tokenización cambiaban.

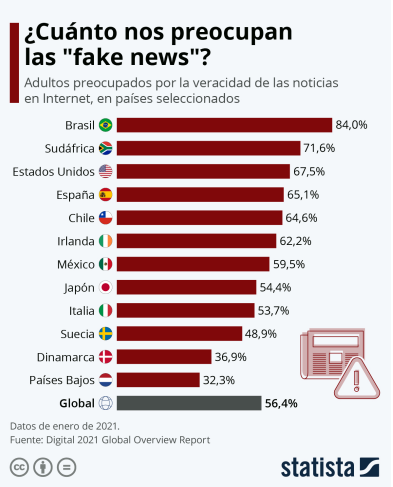
Filtrado fake news

Las **"Fake News"** representan uno de los fenómenos con mayor impacto a nivel de telecomunicaciones y noticias a nivel internacional, representan un grave riesgo tanto para el ciudadano común como para estructuras y modelos que crecen y aprenden a partir de datos e información. Según una encuesta realizada por Digital 21, más de la mitad de adultos a nivel mundial muestran al menos una significativa preocupación por las fake news, ya que estas pueden contener mensajes maliciosos, perjudiciales o erróneos sobre una figura, empresa o sector de forma que genere un impacto negativo tanto en la persona que lee dicha noticia como las nombradas, por ello es tanto pertinente como ético realizar un proceso de filtrado que permita evidenciar la veracidad de las noticias, con el fin de que los gerentes puedan evitar malos entendidos o situaciones incómodas

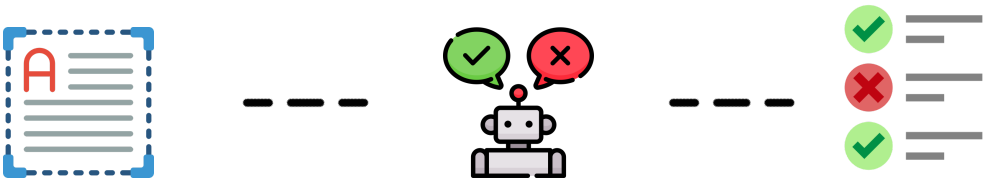
con los clientes del banco, además de ser una práctica ética que responde al desarrollo de sistemas inteligentes establecida por el *marco ético colombiano para el desarrollo de de IA*.

El modelo alBERTa, es una forma optimizada de BERT la cual ha modificado sus hiperparametros eliminando el objetivo de preentrenamiento de la siguiente oración y el entrenamiento con minilotes y tasas de aprendizaje mucho más grandes.

El modelos alBERTa cuenta con resultados muy favorables relacionados a la clasificación de textos contando con un 98% de precisión en modelos de máscara y 96% sin máscara, por otra parte el modelo alBERTa para detección de fake news, cuenta con un excelente resultado del 77,17% de F1 score y una precisión del 77,97, estos son valores bastante positivos teniendo en cuenta la gran dificultad que implica detectar este tipo de noticias ya que cuentan con estructuras gramaticales parecidas a una noticia normal, sin embargo son diferentes en aspectos puntuales, como el uso de excesivas palabras coloquiales, expresiones de odio o divulgación de materiales sensibles para el público en general, estas características son sumamente relevantes y permiten la identificación de estas por medio del modelo.



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 probable de que sea real o fake. Los datos de entrada del modelo deben componerse por la unión del título con el 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”, posteriormente se pondera los score por fragmento de noticia para finalmente unificar los fragmentos con dichos score y obtener la etiqueta real del texto .



Fragmentos de texto preprocesado

Modelo de clasificación de fake news alBERTa

Dataset con las etiquetas de REAL o FAKE

Clasificación

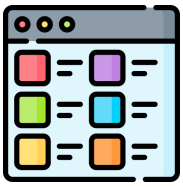
La clasificación de las noticias responde a las necesidades expresadas por los gerentes, donde se evidencian 8 categorías.

Cada categoría cuenta con un nivel de importancia o peso que evidencia la relevancia de esa categoría para los gerentes, esto con el fin de mejorar el repertorio de información que puede tener cada gerente antes de encontrarse con el cliente, la información es clave y permite gestionar estrategias, mejorar la práctica o discurso de interacción con el cliente, generar empatía y fomentar un lazo estable entre empresa-cliente, Bancolombia reconocido por su gran reputación apuesta fuertemente por el cuidado de sus clientes.

Teniendo en cuenta lo anterior el modelo de clasificación seleccionado se basa en BETO, BETO es un modelo BERT entrenado en un gran corpus español. BETO es de un tamaño similar a un BERT-Base y fue entrenado con la técnica de enmascaramiento de palabras completas. BETO permite la clasificación y regresión de palabras en diferentes contextos, en este caso el modelo usado es uno con los hiperparametros modificados de forma que mejora significativamente la precisión en la clasificación en modelos de habla hispana, esto tiene un gran potencial debido a que ya no es necesario generar un proceso de traducción español-inglés para poder utilizar modelos point to point o Blazing text, ya que realizar este tipo de procesos no solo era más dispendioso sino que tiene problemas en la recuperación de información, ya que el español tiene expresiones que dependen mucho del contexto, el cual es fácil de perder a la hora de realizar traducciones literales.

La variante de BETO tuneada, utiliza de base XNLI con un accuracy de 79,9 esto permite obtener un valor probabilístico de las categorías a las cuales se relaciona más el texto introducido, para iniciar el modelo se vuelve a cargar este en una variable por medio del pipeline one_zero_shot.

En el proceso de implementación se pudo encontrar una función extra del modelo que no venía predefinida, ya que se pudo organizar una lista de listas con las palabras de cada categoría y sus sinónimos más cercanos, de esta forma se pueden abordar o relacionar más noticias a un texto mejorando significativamente el resultado de la categorización de forma que se estructuró un csv con los sinónimos (en caso de que requiera ser modificado para ampliar o reducir sinónimos de forma más fluida) estos fueron pasados a vectores que se asignaban a la zona de categorización de la función de clasificación del modelo, junto con los textos, ya que el clasificador no cuenta con un máximo de palabras para el texto que debe clasificar.



Categorías junto a los sinónimos

Modelo de clasificación de categorías BETO

CSV con la categorización de las noticias

Recomendador

La construcción del recomendador parte de la creación de la métrica de aproximación al cliente, esta métrica se crea a partir de la comparación de los términos relacionados al nombre y subsector en el que se desarrolla el cliente con el texto de la notica, estas palabras se usan para construir un vector que mediante la librería de difflib se utiliza la función SequenceMatcher que permite comparar una palabra dentro de un arreglo o texto y seleccionar las palabras que son iguales o cuentan con una mayor similitud, también se contó la cantidad de veces que dicha key se repite en el texto, con ello se calcula la métrica de acercamiento donde weight/total = divide, que sería el valor de la métrica de acercamiento.

Una vez obtenidas las métricas de acercamiento se utiliza la función *describe* para conocer el valor del 75% de los datos en el dataset en la característica divide, de forma que aquellos que tuvieran valores por debajo 0,205% serían noticias que hablan del sector y aquellas por encima, hablan del cliente.

Para la recomendación tendremos en cuenta los peso proporcionados además de dos valores los cuales son la probabilidad de la categoría como un score y el acercamiento al cliente, para ello se unió los dataframe de categorías con el creado con los pesos para así utilizarlos para el cálculo, asignamos los pesos a las categorias como lo sugiere bancolombia, usamos los pesos, el score y el

acercamiento al cliente para calcular el puntaje con el que se hará el recomendador con la siguiente ecuación:
Puntaje de la noticia = Probabilidad de la clase*peso de la clase + probabilidad del el factor de acercamiento

Una vez obtenido el puntaje se ordenan con base en el número NIT del cliente las noticias más relevantes, junto a la participación de la noticia (sector o cliente) obteniendo así el dataset final con las noticias más relevantes por cliente, el sector asociado a esa noticia.

	nombre_equipo	nit	news_id	participacion	categoria	recomendacion
0	Goofy_Goobers	901576411	news11461	Sector	alianza	1
1	Goofy_Goobers	901576411	news79518	Sector	alianza	2
2	Goofy_Goobers	901576411	news96711	Sector	alianza	3
3	Goofy_Goobers	901576411	news79511	Sector	regulación	4
4	Goofy_Goobers	901576411	news98672	Sector	alianza	5

Trabajo a futuro

Para el trabajo a futuro se busca realizar la implementación de los modelos en textos no filtrados, de forma que permiten evidenciar los resultados de la manera más clara y satisfactoria, evitando problemas en la detección de noticias falsas y buscando implementar métricas adicionales al cálculo del puntaje de la noticia como son la probabilidad de que sean noticias reales y la similitud de las noticias de forma que se consigan evitar sesgos y abordar problemas de relevancia en torno a la noticia.

Webgrafía:

hiiamsid/sentence_similarity_spanish_es · Hugging Face
Narrativaai/fake-news-detection-spanish · Hugging Face
Recognai/bert-base-spanish-wwm-cased-xnli · Hugging Face
Dataton (bancolombia.com)
La ciencia confirma que las 'fake news' se extienden más rápido que la verdad (muyinteresante.es)
Search | Statista