1. Contexto

En Colombia, el sector bancario ha sufrido una transformación importante en los últimos años. La entrada de nuevos competidores junto con la mezcla de tecnologías usadas para explotar el campo de la Ciencia de Datos *(Data Science)* en el sector, han facilitado conocer con más particularidad cada uno de los clientes permitiendo crear productos ajustados a las necesidades específicas de cada uno de estos. No obstante, el entender mejor al cliente desde su comportamiento no es la única opción. La información en la internet puede complementar el conocimiento del cliente como su industria, por medio de datos no estructurados tales como noticias o blogs.

Bajo este contexto, Bancolombia como banco de confianza, tiene un reto enorme. Aspira ser ese banquero experto, cercano, que provee soluciones integrales y de largo plazo, y que acompaña al cliente a seguir creciendo y mantener ese liderazgo de mercado. El ofrecer productos hechos a la medida ya no es lo único que los clientes esperan, sino que además aspiran que Bancolombia se anticipe y lo conozca profundamente junto con su entorno, para que los acompañe a alcanzar sus más altos objetivos financieros como no financieros.

No obstante, lograr este objetivo no es sencillo. Actualmente el equipo comercial usa técnicas básicas de investigación en motores de búsqueda, los cuales arrojan miles de noticias sin tener en cuenta si el ranking de los resultados es relevante para el asesor comercial y, si la noticia hace parte de alguna categoría que este esté interesado en leer. Es así, como el equipo comercial debe leer cada una de las noticias, a la expectativa que pueda ayudarles a enriquecer el conocimiento del cliente y su sector para sus futuras reuniones. La ineficiencia de este sistema es evidente sin, además, considerar los desafíos naturales de estas búsquedas como confiabilidad de la fuente, relevancia de la noticia, fuentes múltiples de información entre otros.

Ante dicha ineficiencia, el equipo de Enelpí se ha unido al reto de ayudar a Bancolombia en crear un sistema que permita optimizar el uso de las noticias para un mayor conocimiento del cliente como su sector en varias categorías definidas, facilitando la labor del equipo comercial y apoyándolos en construir relaciones más dinámicas con sus clientes.

2. Introducción

El procesamiento de lenguaje natural (NLP) es una técnica de “machine learning” que brinda a las computadoras la capacidad de manipular e interpretar datos no estructurados a gran escala para comprender el lenguaje humano[1] de forma rápida y eficiente. Sin embargo, la complejidad del lenguaje humano hace que esta técnica este frecuentemente en desarrollo, dado una serie de retos y limitaciones que sigue afrontando[2].

Dentro de ellos, los más importantes para resaltar contienen: ambigüedades, donde se requiere de contexto para comprender una palabra o frase; variedad y dimensionalidad de la data no estructurada, y con ello la demanda de capacidad computacional; desarrollo de algoritmos en diferentes lenguajes; coloquialismos y jergas, con expresiones y terminologías propias de cada

[1] https://aws.amazon.com/es/what-is/nlp/#:~:text=El
[2] https://www.analyticsinsight.net/10-major-challenges-of-using-natural-language-processing/

cultura, y en este proyecto en este especifico, conceptos propios y concretos de cada sector e industria a la cual pertenece un cliente.

Teniendo en cuenta estos desafíos y lo establecido en el proyecto, se ha creado una solución considerando varios modelos usados en NLP, balanceando los pros y contras que cada uno de ellos conlleva, en la solución final. El objetivo de dicha solución es brindarle al equipo comercial noticias confiables, que estén categorizadas y ordenadas por un nivel de importancia, permitiéndoles estar enterados del contexto que rodea al cliente, anticipándose a futuras necesidades con conversaciones de alto valor basadas en la información proveída.

3. Metodología

Basado en los requerimientos del proyecto, la solución se abordó en cuatro partes:

1. Identificación de cliente en la noticia
2. Identificación del sector en la noticia
3. Categorización de la noticia basado en categorías predefinidas
4. Ranqueo de importancia de noticia basado en 5 variables

En cada una de estas partes, se itero de forma constante probando diferentes enfoques con el fin de llegar a una solución holística. En esta sección abordaremos con detalle cada uno de estos enfoques.

3.1 Identificación cliente

Para la identificación del cliente en el texto de la noticia, se usó una metodología ad-hoc que consiste de los siguientes pasos: se crea un diccionario con palabras comunes en los nombres pero que no son útiles para definir la mención a un cliente en especial (i.e cooperativa, grupo, asociación, entre otros). Esto se hace creando n-gramas con los nombres y contando cuantas veces se repiten estos n-gramas, cuando un n-grama tenga más de una mención se señala como que no identifica de manera única a un cliente. Luego, para cada noticia se identifican entidades con NER de la librería Spacy y se compara con los n-gramas del nombre del nit de interés. Las entidades que coincidan con un n-grama del nombre son consultadas en el diccionario pre entrenado, si está en el diccionario se descarta y si no lo está se concluye que el texto sí contiene al cliente. Desde luego, si no hay intersección entre las entidades y los n-gramas se concluye que no hay mención. Un ejemplo de la metodología es considerando el nombre Cooperativa Colanta, la palabra Colanta claramente identifica de manera única al cliente mientras que la palabra Cooperativa no lo hace, si en el texto aparece la palabra cooperativa, el diccionario pre entrenado la descartará como identificación del cliente y requerirá que diga Cooperativa Colanta para concluir que el texto contiene el cliente, por otro lado, si la palabra Colanta aparece por sí sola, esta coincidirá con un n-grama de la frase Cooperativa Colanta y no estará en el diccionario pre-entrenado por lo que se concluirá que el texto sí contiene al cliente.

3.2 Identificación sector

Para la identificación del sector en el texto de la noticia, se empezó usando Latent Dirichlet Allocation (LDA) para encontrar si la noticia contiene el sector que se relaciona con un cliente. LDA es una técnica usada como modelamiento de tópicos que permite clasificar un documento, en este caso una noticia, en un tópico en particular dado un corpus. Es llamado “Dirichlet” dado que uno

usa la distribución Dirichlet para encontrar los tópicos en cada uno de los documentos por medio de palabras claves. Es importante mencionar que el tratamiento del texto de las noticias fue importante para este algoritmo removiendo caracteres especiales, palabras “vacías” y números.

LDA permite el paso de múltiples parámetros para mejorar el algoritmo, uno de ellos conocido como semillas. Las semillas son palabras claves que el usuario usa para guiar al algoritmo hacia ciertos temas. En este caso particular, las semillas que se usaron fueron las palabras encontradas en las columnas división, grupo, clase y subsector. Para las noticias que contienen palabras muy cercanas a las semillas, el algoritmo funciona muy bien. Sin embargo, si la noticia tiene un texto que no se relaciona directamente a las semillas, tiene limitaciones de hacer relaciones indirectas del texto con las semillas.

Para sobrepasar esta limitación, se decidió usar un algoritmo de red neuronal pre-entrenado “Sentence Transformer” *paraphrase-multilingual-mpnet-base-v2*, el cual permite crear palabras “embebidas” en un vector de alta dimensionalidad – 512 dimensiones – estableciendo relaciones no directas del texto con las semillas. En este enfoque, se decidió comparar cada una de las columnas (división, grupo, clase y subsector) con los párrafos de cada artículo usando la “similitud de coseno ”, creando una matriz m x n; m siendo las columnas usadas como semillas y n siendo el número de párrafos en la noticia. Para cada medición de “similitud de coseno” se aplicaron dos variaciones para definir si la noticia contenía el sector:

1. La matriz debe haber contenido al menos 5 frases con al menos 0.4 de similitud con alguna de las columnas
2. La matriz debe haber contenido al menos 3 frases con al menos 0.5 de similitud con alguna de las columnas

Como último, para definir si la noticia contenía el sector, se creó un sistema de votación considerando el resultado del LDA y el “Sentence Transformer” con sus dos variaciones. La votación funciona de tal forma que, si 2 de 3 de los resultados contiene el sector, la noticia es marcada como que contiene al sector del cual el cliente hace parte.

3.3 Categorización de la noticia

Para la categorización de la noticia en el texto de la noticia, se empezó usando *LDA* para clasificar la noticia en las categorías definidas por el equipo de Bancolombia. La forma en que se abordó este paso fue creando un corpus con todos los textos de las noticias, previamente removiendo caracteres especiales, palabras “vacías” y números. Luego de eso, se corrió el modelo LDA para empezar a investigar cuales eran los tópicos que eran claros dentro del corpus. Después de haber identificado algunos tópicos en la primera iteración, se volvía a iterar usando las palabras claves arrojadas por tópico para guiar un poco más al algoritmo hacia palabras deseadas dentro de los tópicos. Se hicieron tres iteraciones para encontrar las semillas claves, dado que en la cuarta y quinta iteración la variación de las palabras no era significativa.

Luego de guiar el LDA con las semillas de la tercera iteración, el algoritmo LDA funcionó muy bien para categorías bastante bien definidas dentro de los documentos tales como Macroeconomía, Regulaciones y Sostenibilidad. Sin embargo, la precisión para otras categorías no era la mejor, y el

equipo decidió probar el algoritmo de KeyBERT[3] utilizando un modelo red neuronal pre-entrenado llamado *paraphrase-xlm-r-multilingual-v1.* KeyBERT usa la técnica base de extracción BERT, la cual transforma las palabras del documento en vectores de alta dimensionalidad (512 dimensiones) para aplicarles la similitud de coseno, encontrando y extrayendo los n-gramas en un documento lo más similar posible al documento mismo. KeyBERT guiado es una variación del KeyBERT simple, la cual fuerza la transformación de los vectores hacia semillas (palabras) proporcionadas como parámetro, con el fin de encontrar n-gramas que se relacionen con las semillas proveídas.

Una de las grandes ventajas de KeyBERT, es que permite el uso de modelos pre-entrenados para la transformación de las palabras en los vectores de alta dimensionalidad que más se ajuste al caso en particular. En este ejercicio, decidimos usar *"paraphrase-xlm-r-multilingual-v1"*[4] dado que requeríamos un modelo que nos permitiría transformar las palabras en español al vector de forma rápida y eficiente, sin necesidad de entrenar un modelo dada la complejidad de la clasificación. Además, debíamos tener en cuenta la limitante tanto en la cantidad de datos proveídos en las noticias como el poder computacional necesario para entrenar un modelo desde ceros para obtener resultados sobresalientes.

El algoritmo de KeyBERT fue usado con dos enfoques diferentes: El primer enfoque - KeyBERT simple - fue usado para fue extraer los 5 n-gramas más relevantes dándole la libertad al algoritmo de escoger las palabras más relevantes de la noticia sin ninguna guía. El segundo enfoque - KeyBERT guiado - fue usado para obtener los 5 n-gramas más importantes usando palabras claves para forzar el algoritmo a dirigirse a las palabras importantes deseadas. A los resultados de los n-gramas arrojados por KeyBERT junto con las categorías, les empleamos una transformación a vectores de alta dimensionalidad aplicando el modelo pre-entrenado *"google/universal-sentence-encoder-multilingual-large/3"* para luego calcular la semejanza entre la categoría y los n-gramas del documento por medio de la técnica "similitud de coseno". Cada n-grama conseguido por KeyBERT fue comparado con una de las categorías en el espacio de alta dimensionalidad, y cada uno de estos resultados fueron sumados por categoría. La categoría con mayor valor en su suma fue escogida como la categoría dominante de la noticia.

Como último enfoque, y dado los buenos resultados que obtuvimos en la etapa anterior, decidimos agregarle a la categorización un "Sentence Transfomer" (similar al usando en la identificación del sector), pero esta vez aplicado a cada uno de los párrafos de las noticias. Así, se creó una matriz de n x m calculando la "similitud de coseno" entre categorías y párrafos; m siendo cada categoría en una columna y n siendo el número de párrafos en la noticia. La matriz fue reducida a una categoría única por noticia, aplicando un promedio de cada categoría a través de cada uno de los párrafos y extrayendo la categoría de mayor valor.

Con el fin de tener una mayor certeza si los resultados de estos algoritmos estaban funcionando, y cuál era el que mejor entre los cuatro, el equipo decidió clasificar 300 noticias manualmente. Estas 300 noticias fueron escogidas basadas en el criterio que cada una de las variaciones de los algoritmos daba como resultado una clasificación diferente. Cada uno de los integrantes leyó las noticias y le asignó una categoría la cual consideraba que la noticia pertenecía. Con este ejercicio, nos dimos cuenta que habían noticias con mucha ambigüedad. A lo que nos referimos con esto, es

[3] Quickstart - KeyBERT (maartengr.github.io)

[4] https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1

por ejemplo que había noticias que hablaban de innovación para crear un sistema sostenible, y para un algoritmo era claro que la noticia podría ser clasificada como innovación, pero para los otros podría llegar a ser considerada sostenibilidad, y en algunos casos podría ser clasificada como otra - posiblemente por la posibilidad de ser categorizada como Tecnología.

Basado en esto, luego de explorar varias opciones, el equipo decidió expandir a las categorías un poco más. El equipo creo categorías como: Deporte, Política, Infraestructura, Tecnología, Cultura, entre otras. Con estas nuevas categorías, los algoritmos tuvieron una mejora significativa, y los diferentes algoritmos convergían mucho mejor hacia una categoría definida. Al comparar una vez más la clasificación manual vs la clasificación de los algoritmos, el modelo con mejor resultado fue el "Sentence Transformer" usado sobre párrafos. A pesar de esto, el equipo comprendió que la mejor solución era combinar los resultados de los diferentes algoritmos bajo un sistema de votación. El desempeño de este sistema de votación es superior al desempeño de cualquiera de los 4 modelos individuales por al menos 20% en precisión.

El sistema de votación funciona de tal manera que, si los 4 algoritmos tienen diferentes resultados o dos algoritmos tiene la misma categoría que los otros dos restantes, el valor predeterminado en la categoría es el resultado del modelo "Sentence Transformer" enfocado en los párrafos dado que su rendimiento sobresale entre los 4. Si dos o más de los 4 algoritmos tienen un valor similar entonces se determina la categoría por la frecuencia que se encuentre a través de los algoritmos.

3.3 Ranqueo de noticias

Para el ranqueo de noticias se manejaron 5 variables que se consideraron importantes:

1. Contiene al cliente: Si una noticia contiene al cliente, se le da un peso de 1. De lo contrario se le da un peso de 0 (cero).
2. Contiene al sector: Si la noticia contiene el sector, se le da un peso de 1. De lo contrario se le da un peso de 0 (cero)
3. Peso por categoría: Los valores de las categorías proveídas por el equipo comercial fueron estandarizadas sobre 1. Es importante resaltar que se decidió darle un peso mucho más importante a la categoría "Reputación" dado que consideramos que para Bancolombia debe ser una prioridad construir relaciones con clientes alienados a la misma imagen que Bancolombia quiere proyectar con sus demás clientes. Por ejemplo, para un valor no estandarizado de 2.1 en la categoría "Reputación", estandarizado es 1. Para la categoría Macroeconomía con un valor no estandarizado de 2, estandarizado es 0.9523. Así sucesivamente se hizo con todas las categorías.
4. Fuente de la noticia: Las fuentes fueron extraídas usando los URLs de las noticias. Con ellas se hizo una distribución de la frecuencia en que sucedían por categoría. Las fuentes con mayor frecuencia tienen un puntaje mucho mayor. Esto se hizo teniendo en cuenta que algunas fuentes se especializan más en unos temas más que otros (Silla Vacía – Política vs. Semana – Macroeconomía vs. Marca.com - Deportes)
5. Longitud de la noticia: Luego de hacerle un tratamiento de limpieza a la noticia, se hizo una distribución de la longitud del texto de la noticia. Luego de un análisis, a las noticias con una longitud menor a 31 palabras, se le dio un peso de 0. Por otro lado, las noticias con mayor longitud, se les dio el peso más alto considerando de la distribución de las longitudes por categoría.

Usando estas cinco variables, se hizo una sumatoria para obtener un valor final como ranqueo. Luego, se hizo uso un ranqueo denso para clasificar las noticias de 1 a N por cliente ordenado de mayor a menor puntaje sobre las 5 variables.

6. Recomendaciones

• Almacenar el conocimiento previo del cliente: La percepción de la importancia de una noticia para conocer un cliente está condicionada al conocimiento previo que el asesor tiene del mismo. Por ejemplo, si un cliente está inmerso en una investigación judicial, una noticia relacionada puede parecer importante para quien no lo sepa, pero irrelevante para quién ya conocía esta investigación. Sin embargo, una noticia que ahonde en nuevos hallazgos de la investigación o una conclusión de esta puede ser de verdad relevante para el asesor. Debido a esto, recomendamos que el sistema de recomendación implemente un módulo de conocimiento previo y que este sea sumado al modelo de recomendación de manera que las noticias con información no vistas antes tengan mayor prioridad, como que noticias con información redundante sean descartadas. En nuestra implementación asumimos que el asesor tiene cero previo conocimiento del cliente.

• Sistema de resumen de noticias: Es evidente en el resultado del ranqueo de noticias que el sistema puede ser una parte fundamental de apoyo para la labor del equipo comercial. Sin embargo, el tiempo que puede llegar a tomar en leer las primeras M noticias por N clientes puede llegar a ser substancial. Se recomienda en trabajar en un sistema de resumen de la noticia, extrayendo las frases más relevantes permitiendo un resumen global de la noticia, creando un sistema de resumen de noticias que optimice la labor de lectura para los asesores comerciales.

• "*Hard Mining y Data Mining*" para datos no balanceados[5]: Los modelos implementados en este proyecto pueden ser mejorados mediante el uso de estas dos técnicas que usan un etiquetado secuencial para que los modelos sean capaces de identificar casos difíciles y casos con poca representatividad en la base de datos. Nuestra implementación de los modelos de predicción de categoría usa sutilmente uno de estos algoritmos, pero creemos que se puede mejorar dado que los recursos de etiquetado tendrían un costo razonable para Bancolombia con una mejora sustancial en los modelos.

• Uso de API's de noticias: Durante el proyecto, el equipo de Enelpí se percató que había noticias con una longitud anómala y decidió obviarlas dado la cantidad de recursos computacionales que requerían. Además, se evidencio que el "scraper" usado para recolectar la información, capturo data redúndate tal como "También le interesaría leer", "Inicio de sesión", entre otras cosas. Se recomienda usar un API que permita jalar información limpia y confiable en un proceso ya en producción.

[5] https://arxiv.org/pdf/1604.03540.pdf & https://www3.nd.edu/~dial/publications/chawla2005data.pdf