Trabajo de Fin de Máster
Máster Técnicas Estadísticas
Fecha: 03-06-2024

Mejora de la Eficiencia Bancaria mediante NLP: Segmentación de llamadas transcritas para la identificación de Productos Comerciales

Marcos Gómez Rodríguez

Este documento resume el proyecto del Trabajo Fin de Máster *Mejora de la Eficiencia Bancaria mediante NLP: Segmentación de llamadas transcritas para la identificación de Productos Comerciales* dirigido por don Javier Tarrío Saavedra, Profesor Titular de la Universidade da Coruña, don Salvador Naya Fernández, Catedrático de la Universidade da Coruña y doña María González Estévez, Mánager del equipo de Analítica Predictiva de ABANCA. Por motivos de confidencialidad no es posible la publicación de la memoria completa.

Resumen

Este proyecto se realizó en el área de Inteligencia de Clientes de la entidad ABANCA Corporación Bancaria S.A que se dedica a la inteligencia de negocio o Business Intelligence (BI). Su principal objetivo es el uso de herramientas analíticas para transformar los datos en conocimiento de negocio aportando valor en el proceso estratégico de toma de decisiones de la compañía. En particular, este área se especializa en el proceso de recopilación, análisis de los datos disponibles, desarrollo de modelos analíticos, creación de cuadros de mando y seguimiento de KPIs (*Key Performance Indicators*), con el objetivo de entender las necesidades financieras de los clientes para poder adelantarse a las mismas mediante una orquestación de acciones comerciales personalizadas a cada uno.

Desde el 2017, ABANCA ha impulsado un modelo de relación omnicanal con los clientes en el que además de las formas tradicionales de comunicación, como la atención en oficinas, se le da importancia a canales digitales como la Banca Móvil o la Banca Electrónica. En este contexto se crea ABANCA Conecta, una unidad de gestión personalizada a distancia, que nace, según se explica en la Intranet de ABANCA, de la necesidad de dar servicio a un segmento de clientes que prefieren no desplazarse, pero que sin embargo demandan un asesoramiento comercial a distancia. Este servicio de ABANCA, complementa al resto de

la red comercial de la entidad, ofreciendo asesoramiento al cliente para realizar operaciones comerciales o gestiones de manera presencial, en oficinas o desde canales digitales.

Desde Inteligencia de Clientes se ha colaborado con Conecta poniendo a disposición herramientas analíticas como modelos de propensión de contratación de productos financieros para aumentar la eficacia de las oportunidades comerciales de los gestores priorizando en su cartera a los clientes con una probabilidad elevada de estar interesados. Esto no solo es de utilidad para los propios gestores, que ayuda en incrementar los éxitos de la contactabilidad y por tanto mejoran su productividad, si no que también se mejora la experiencia del cliente, ya que reciben una atención más alineada con sus intereses.

Uno de los principales objetivos del departamento es aprovechar la información verídica procedente de los contactos entre gestor y cliente para enriquecer la información interna. Entre ella destaca las transcripciones de algunas llamadas entre gestores y clientes a través del servicio ABANCA Concecta. Destacar que cumpliendo las leyes de privacidad establecidas, todos los clientes que tienen el servicio de Conecta previamente se les comunica y deben aceptar el uso de los conversaciones con su gestor.

Debido a que estas llamadas no están etiquetadas, desde Inteligencia de Clientes se propuso el uso de modelos de lenguaje natural para extraer el contenido de las llamadas de manera automática y no supervisada. Este trabajo tiene como finalidad el estudio de los modelos de temas para su aplicación en las transcripciones de las llamadas de ABANCA Conecta y su posterior clasificación de la información para mejorar y enriquecer los procesos analíticos del departamento.

En Axelborn & Berggren (2023) se utilizaron dos modelos de temas, LDA y BERTopic, para el análisis de temas de llamadas telefónicas en un contact center.

En Vandevoort et al. (2023), se estudia la utilidad de los modelos de temas en varios casos de uso dentro de la EFSA (European Food Safety Agency). Estos casos incluyen agrupar comentarios similares recibidos para facilitar la tarea de responderlos, así como explorar un conjunto de documentos que contienen la palabra "Beeswax". La dificultad que afrontaban es que, al filtrar por esta palabra, aparecen documentos y artículos de otras áreas no relacionadas con la alimentación. Por ello, utilizan modelos de temas para filtrar específicamente los artículos alimentarios, ayudando a los expertos a localizar los textos relevantes sin tener que revisar cada documento individualmente. Utilizan técnicas de modelado de temas clásicas como LDA y NMF, así como técnicas recientes basadas en embeddings de redes neuronales como BERTopic y Top2Vec. En Egger & Yu (2022), se comparan estos modelos de temas en el contexto del análisis de datos procedentes de Twitter.

En este trabajo se aplican los modelos de temas LDA, NMF y BERTopic y Top2Vec para la identificación no supervisada de los temas de las llamadas telefónicas de ABANCA Conecta con el objetivo de extraer información potencial de las transcripciones, que pueda servir para mejorar la eficiencia de los modelos predictivos con los que se trabaja en el departamento.

Para ello se hace un estudio detallado de los conceptos teóricos que se usan durante el trabajo, que se dividen en métodos para obtener representaciones numéricas de textos y modelos de temas.

La importancia de obtener representaciones numéricas de los textos radica en de esta manera es posible utilizar aplicarles modelos de Aprendizaje Estadístico que solo pueden usarse si se trabaja con números o distancias. En el trabajo se ha distinguido dos tipo de métodos para obtener representaciones numéricas de los textos.

- **Basadas en frecuencias de palabras.** Una forma clásica de representar numéricamente un texto es a través del conjunto de palabras que lo forma. De esta forma se puede obtener una matriz de tantas columnas como el tamaño del vocabulario, en la que cada documento se representa como un vector cuyas componentes son el número de veces que aparece las palabras en él. Esto se conoce como bolsa de palabras o *Bag of Words*. Una de las modificaciones habituales de este método es lo que se conoce como Tf-Idf, *Term frecuency - Inverse Document Frecuency*, que consiste dividir la frecuencia de aparición de cada palabra en los documentos entre la proporción de documentos en los que aparece dicho término. Estos métodos no tienen en cuenta el orden de aparición de las palabras. Otra desventaja de estos modelos es que no son capaces de identificar sinónimos ni palabras con varios significados.
- **Basadas en redes neuronales.** Entre estos métodos destacan el modelo doc2vec de Mikolov et al. (2013) y modelos más recientes como el *multilingual-e5-base* (Wang et al., 2024), entrenado a partir de la arquitectura BERT (Devlin et al., 2019) sobre un corpus masivo de textos de internet. El modelo *multilingual-e5-base* es capaz de entender el contexto, pudiendo lidiar con sinónimos. Además, debido a que fue entrenado con textos de varias lenguas, es capar de reconocer similitud de textos aunque estos no estén en el mismo idioma.

En cuanto a los modelos de temas, se han estudiado los modelos LDA, NMF, BERTopic y Top2Vec, como se hace en Egger & Yu (2022).

- **Latent Dirichlet Allocation.** Latent Dirichlet Allocation (LDA) es un modelo de temas propuesto por Blei et al. (2003), que conceptualiza los documentos como

colecciones de temas. En LDA, se supone que cada documento es una mezcla de diversos temas, y cada tema, a su vez, es una distribución sobre un vocabulario de palabras. Este modelo trata de capturar la generación de palabras en los documentos a través de procesos estocásticos basados en dos distribuciones estadísticas: la Multinomial y la Dirichlet.

En el modelo LDA, se selecciona primero una distribución discreta de los temas para un documento dado. Para cada palabra nueva en el documento, se elige un tema de manera aleatoria, siguiendo la distribución de temas. Luego, cada tema, que es una distribución sobre el espacio de palabras, genera la palabra correspondiente. Esta metodología no considera que el orden en el que aparecen las palabras aporte información, como ocurría con la bolsa de palabras.

- **Non Negative Matrix Factorization.** La factorización matricial no negativa (NMF) es una técnica para descomponer una matriz no negativa con valores reales V en otras dos W, H también no negativas, de forma que $V \approx WH$ (Lee & Seung, 1999). Este método, al igual que el Análisis de Componentes Principales, puede ser usar para reducir la dimensionalidad de un conjunto de datos.

  En Lee & Seung (1999) se explica que PCA y NMF se basan en obtener una descomposición de unos datos en formato matricial V, de dimensión $n \times m$ de forma que

  $$V_{ij} \approx (WH)_{ij} = \sum_{a=1}^{k} W_{ia} H_{aj},$$

  diferenciándose ambos métodos en las restricciones que se imponen en las matrices W y H.

  En el método de PCA, las columnas de W deben ser ortonormales y las filas de H deben ser ortonormales entre sí, exrpesando V como combinación lineal de una base formada por autovalores, las r columnas de W. El método de NMF no utiliza la hipótesis de ortonormalidad pero impone que los elementos de ambas matrices sean no negativos. La utilidad de esta restricción es ilustrada en Lee & Seung (1999) con un ejemplo en el que se usan datos de imágenes procedentes de caras. La información de las imágenes es convertida en en vectores con una alta dimensión y se utilizan ambos métodos para reducirla. Las bases obtenidas al aplicar el método de PCA, eran poco interpretables, ya que al permitir combinaciones tanto positivas como negativas, las caras estaban generadas con imágenes de características complejas que a menudo se cancelaban entre ellas. Exigiendo no negatividad, las bases obtenidas resultaron partes de la cara, como ojos, orejas, bocas, etc. Usando el método de

NMF se obtuvo por tanto una descomposición de las caras como suma de partes de estas.

Esta misma idea se puede trasladar al ámbito de datos de texto. Si se aplica la descomposición NMF a la matriz de frecuencias, que es no negativa, se puede obtener una aproximación de esta formada por combinaciones positivas de vectores de palabras (que también serán positivos) y que representarán los temas del texto, de forma análoga a las partes de las caras obtenidas en Lee & Seung (1999).

- **Top2Vec.** En Angelov (2020) se introduce esta técnica de modelado de temas a partir de las representaciones numéricas del modelo Doc2Vec. En dicho artículo se propone el esquema a seguir que después seguirá el modelo BERTopic (Grootendorst, 2022) , dividiendo el modelado de temas en varias etapas.

  La primera parte consiste en extraer las representaciones numéricas de los textos mediante redes neuronales, para seguidamente realizar una reducción de la dimensionalidad. Es importe reducir la dimensionalidad de los datos para mejorar el rendimiento de la clusterización, al trabajar sobre un conjunto más denso. Una vez obtenida una representación vectorial de los textos con menor dimensión, se aplica un modelo de clusterización sobre estos datos. Una vez realizada la clusterización, se extraen las palabras más relevantes de cada uno de los clústeres. Por último, se extraen las palabras más relevantes de cada tema, que en Angelov (2020) se obtenían como las palabras más cercanas a cada uno de los centroides de los clústeres. En Angelov (2020) se propone reducir la dimensionalidad usando UMAP (*Uniform Manifold Approximation and Projection*) y hacer la clusterizacón mediante HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*). La representación numérica de los textos, conocida como embedding, se obtiene mediante doc2vec.

- **BERTopic.** Sigue el mismo esquema que Top2Vec pero utilizando embeddings preentrenados como los basados en BERT. Además, en (Grootendorst, 2022) se propone el uso de c-Tf-Idf (*class based Tf-Idf*) para obtener las palabras más relevantes de cada tema. Esta técnica consiste en concatenar los documentos de cada tema y obtener el Tf-Idf sobre dicho conjunto.

El siguiente paso fue un análisis de los datos disponibles, junto a un preprocesado de los textos. Este preprocesado consistió en eliminar nombres, palabras frecuentes, caracteres especiales y signos de puntuación, así como un *stemming* que devuelve las palabras a su forma raíz, de forma que distintas conjugaciones de los verbos se asignen a la misma palabra. Esto es de especial relevancia al trabajar con LDA y NMF, ya que se costruyen

a partir de la matriz de frecuencias de palabras. Para los modelos BERTopic y Top2Vec solamente se pasó el texto a minúsculas y se eliminaron los nombres, debido a que estos modelos funcionan bien con poco procesamiento de los textos. Por último, se filtraron las llamadas en las que aparece un mensaje del contestador automático, ya que esto supuso una mejora en los modelos ajustados.

Para elegir los parámetros de los modelos se ha utilizado la coherencia de temas (Röder et al., 2015), de forma que los finalmente seleccionados eran los que maximizaban dicha métrica. En el caso de BERTopic se encontró que el máximo de coherencia se alcanzaba considerando solo 2 temas, lo que no era de utilidad para el objetivo del trabajo, por lo que en su ajuste se penalizó la coherencia por la cantidad de temas obtenidos.

En cuanto al software empleado para la construcción de los modelos, se utilizó R para el procesamiento de los datos, en concreto el paquete `tm`. Para ajustar los modelos de temas se han usado los paquetes de Python `gensim` para LDA y NMF, `BERtopic` para el ajuste del modelo BERTopic y `Top2Vec` para el modelo Top2Vec. También se probó la implementación de NMF de `Sklearn`, obteniendo mejores resultados que la de `gensim`. Para obtener la coherencia de temas se utilizó el paquete `gensim`.

En cuanto a los resultados obtendidos, todos los modelos fueron capaces de reconoces temas asociados a distintos productos de la entidad como seguros o hipotecas. BERtopic es más rápido que Top2Vec al usar un modelo pre-entrenado, pero Top2Vec, al estar entrenado sobre las transcripciones, fue capaz de identificar vocabulario específico, como un Tema relacionado con descubiertos y números rojos, que es el nombre que recibe el suceso de quedarse sin saldo en una cuenta bancaria. Ambos modelos son capaces de detectar temas con más detalle que LDA y NMF, gracias a que la clusterización jerárquica de HDBSCAN permite obtener un árbol con temas y subtemas de los mismos. NMF consiguió mejores resultados que LDA, y parece más apropiado para extraer características de los textos gracias a su propiedad de reducción de dimensión de la matriz de frecuencias de palabras. Además, la implementación del modelo NMF de `Sklearn` consiguió resultados mucho mejores que la de `gensim` con menor tiempo de computación.

Referencias

Angelov, D. (2020). Top2vec: Distributed representations of topics.

Axelborn, H. & Berggren, J. (2023). *Topic Modeling for Customer Insights A Comparative*

Analysis of LDA and BERTopic in Categorizing Customer Calls. Umea University, Sweden.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null), 993–022.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Egger, R. & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, 7. https://doi.org/10.3389/fsoc.2022.886498

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Lee, D. D. & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791. https://doi.org/10.1038/44565

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. https://doi.org/10.48550/ARXIV.1310.4546

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM 2015. https://doi.org/10.1145/2684822.2685324

Vandevoort, B., Bex, G., Crevecoeur, J., & Neven, F. (2023). Topic modelling and text classification models for applications within efsa. *EFSA Supporting Publications*, 20(8). https://doi.org/10.2903/sp.efsa.2023.EN-8212

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual e5 text embeddings: A technical report. https://doi.org/10.48550/ARXIV.2402.05672