PROYECTO FINAL

Análisis de Opiniones Negativas para Priorizar Atención al Cliente

Procesamiento de Lenguaje Natural

Maestría en Cienciade Datos

Universidad de Sonora

Alumnos:
Estephania Pivac Alcaraz
Victor Manuel Minjares Neriz

Profesor:
Gerardo Mauricio Toledo Acosta

03 de Diciembre de 2024

Contenido

Contexto

La empresa, dedicada a la venta de alimentos por internet, enfrenta el desafío de garantizar la satisfacción del cliente, priorizando la atención a aquellos con experiencias negativas. Se cuenta con un conjunto de reseñas que incluye un texto y una calificación en una escala de 1 a 5 (1 siendo la más baja y 5 la más alta).
El objetivo principal es desarrollar un sistema que:

1. Identifique automáticamente las reseñas más negativas (calificaciones de 1 y 2).
2. Analice los temas predominantes en estas opiniones negativas para responder: ¿Cuáles son las quejas más comunes?

Este informe detalla los métodos utilizados en cada etapa, desde el preprocesamiento hasta la evaluación, los detalles técnicos de los modelos empleados y los resultados obtenidos.

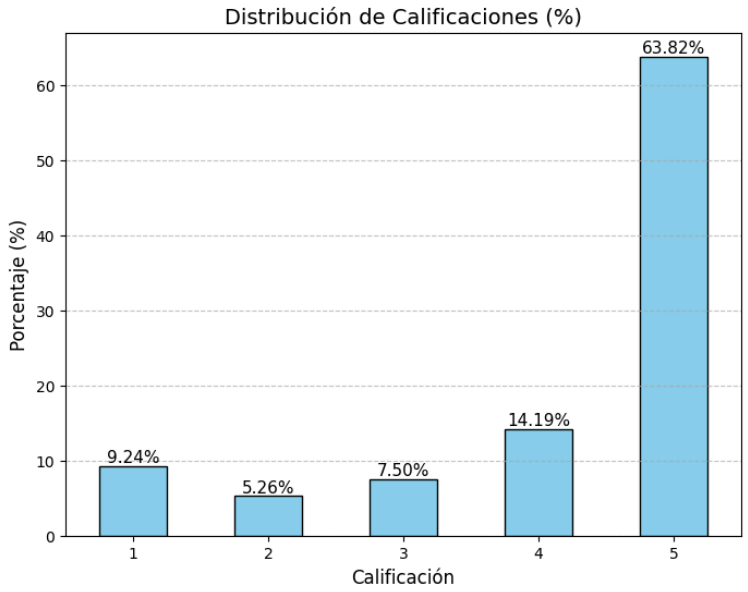
Preprocesamiento de Datos

En el análisis exploratorio, se identificó un desequilibrio en las clases: solo el 14% de las reseñas son negativas. Para abordar este desequilibrio, se utilizaron las siguientes estrategias:

- Estratificación de los datos al dividirlos en conjuntos de entrenamiento y prueba.
- Uso de una métrica de evaluación que considere el desbalance de clases.
- En el caso del modelo DistilBert se realizó un submuestreo asegurando el balance de clases para hacer el entrenamiento, lo cual abordaremos con detalle más adelante.

Se realizaron las siguientes tareas de preprocesamiento:

- Limpieza de datos:
    - Eliminación de registros duplicados.
    - Exploración de valores nulos y de la distribución de las calificaciones en la muestra.



    - Consolidación de las columnas Summary y Text en una única columna para conservar toda la información.
- Limpieza enfocada en PLN:
    - Eliminación de *stopwords*, caracteres no alfanuméricos, etiquetas HTML dentro de < >, y espacios blancos múltiples.
    - Tokenización.
    - Transformación de las calificaciones en una clasificación binaria: reseñas negativas (1) y positivas (0).

No se aplicaron técnicas de lematización ni *stemming* para evitar pérdida de información y optimizar el tiempo de cómputo.

Modelos de Clasificación de Sentimientos

1. Naive Bayes con Vectorización TF-IDF

Justificación:
Se eligió este modelo debido a su simplicidad, rapidez y desempeño demostrado en tareas de análisis de texto.

Hiperparámetros:

- max_features: Cantidad máxima de características en la vectorización.

- class_prior: Distribución a priori de las clases.
  Estos valores se determinaron mediante experimentación a partir de la razón calculada mediante el desbalance de las clases, optimizando la métrica balanced_accuracy y considerando el tiempo de cómputo.

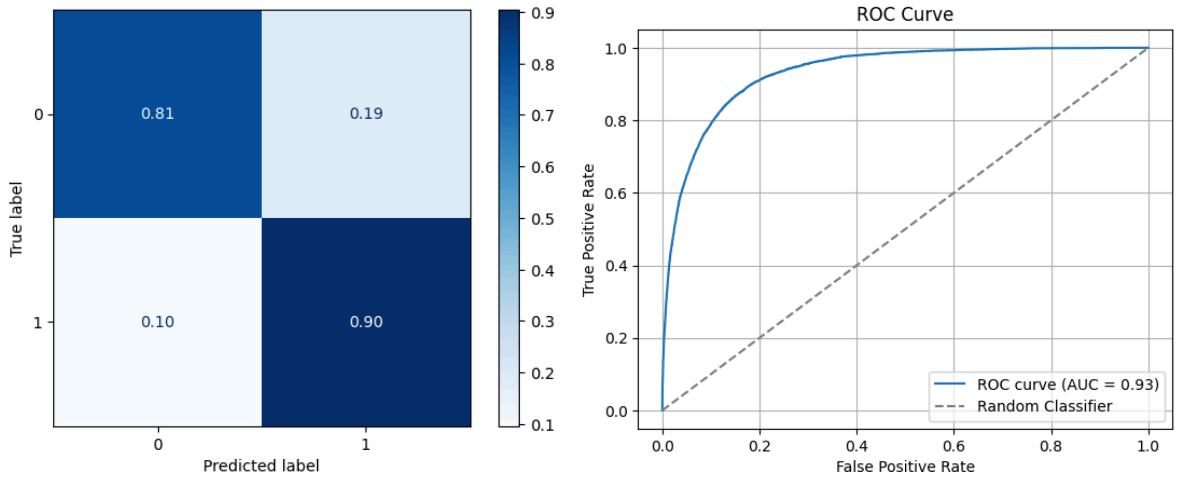
Métricas de Evaluación:

- balanced_accuracy: Ideal para conjuntos desbalanceados.
- Precision: Proporción de verdaderos positivos entre todas las predicciones positivas. En nuestro caso, evalúa que tan confiable son las predicciones de sentimiento negativo entre los reviews, que es justo el objetivo.
- Recall: Proporción de verdaderos positivos sobre todos los positivos reales. Esta evalúa la capacidad que tiene el modelo para capturar todos los casos positivos (en nuestro caso, los reviews negativos).
- F1-Score: Por ser la media armónica de precision y recall, mide un equilibrio entre ambas métricas. Es útil cuando manejamos dataset desbalanceados como en nuestro caso.
- Curva ROC: Indicador del desempeño general del modelo.

Resultados:
El modelo Naive Bayes con TF-IDF ofreció resultados satisfactorios en términos de precisión y rapidez, superando las expectativas considerando su simplicidad.

	precision	recall	f1-score	support
0	0.98	0.81	0.89	52797
1	0.45	0.90	0.60	9027
accuracy			0.82	61824
macro avg	0.71	0.86	0.74	61824
weighted avg	0.90	0.82	0.85	61824

Balanced Accuracy: 0.86



2. DistilBERT

Justificación: DistilBERT, una versión ligera de BERT (Bidirectional Encoder Representations from Transformers), se consideró un modelo adecuado para la tarea de clasificación de sentimientos por su equilibrio entre eficiencia computacional (por ser más pequeño que otros modelos basados en BERT) y por su rendimiento. Adicionalmente, DistilBERT aprovecha el aprendizaje transferido de BERT y nos permite realizar fine-tuning con nuestro dataset específico del proyecto.

Metodología:

- Se realizó un submuestreo de los datos de entrenamiento asegurando el balance de ambas clases (6000 datos en total). Esto con la finalidad de reducir tiempos de entrenamiento. Se separó este conjunto balanceado en train y validation en una proporción de 50%/50%, esto porque queremos entrenar con pocos datos debido a los recursos limitados
- Se realizó una tokenización y padding dinámico para asegurar que todas las secuencias tengan la misma longitud sin perder información importante.
- Se optó por 5 epochs principalmente por el riesgo de sobreajuste. Se utilizó el hiperparámetro de warmup_steps=500 para estabilizar el entrenamiento al comienzo y se utilizó weight_decay=0.01 para regularizar el modelo (evitar un sobreajuste).
- Se utilizó Trainer para facilitar el entrenamiento y optimizar automáticamente los parámetros del modelo y el manejo del GPU. Además de que permite una evaluación contínua en cada epoch.
- El modelo optimiza por default utilizando Binary Cross-Entropy-Loss ya que se trata de un problema de clasificación binaria.
- El modelo se entrenó con la submuestra del conjunto de training y se evaluó utilizando todo el conjunto test, asegurándose de realizar el mismo preprocesamiento para obtener los resultados.

Métricas de Evaluación:

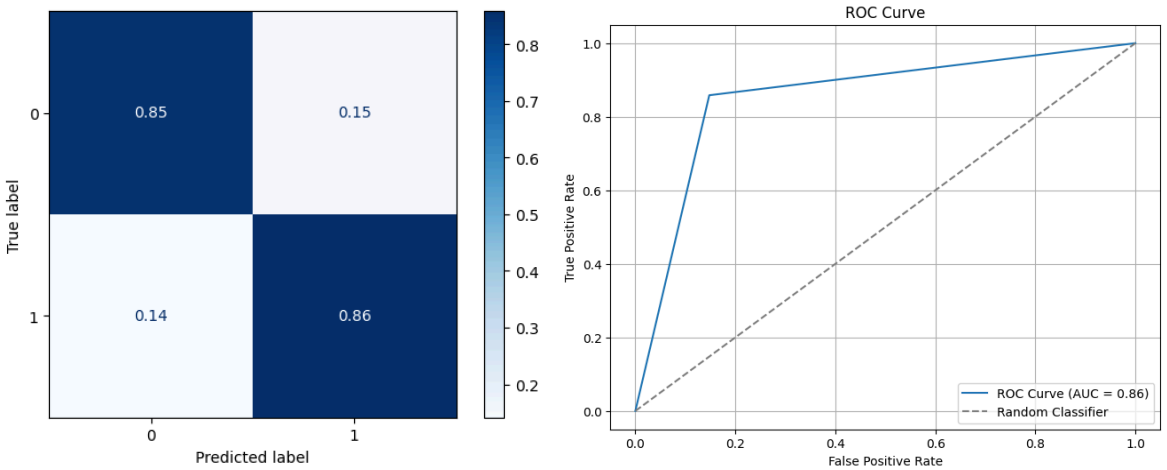
- Balanced Accuracy: Promedio de recall por clase, asegurando que ambas clases contribuyan de igual manera al resultado final, lo cual es útil cuando tratamos datos desbalanceados (como nuestro caso).
- Precision: Proporción de verdaderos positivos entre todas las predicciones positivas. En nuestro caso, evalúa que tan confiable son las predicciones de sentimiento negativo entre los reviews, que es justo el objetivo.

- Recall: Proporción de verdaderos positivos sobre todos los positivos reales. Esta evalúa la capacidad que tiene el modelo para capturar todos los casos positivos (en nuestro caso, los reviews negativos).
- F1-Score: Por ser la media armónica de precision y recall, mide un equilibrio entre ambas métricas. Es útil cuando manejamos dataset desbalanceados como en nuestro caso.
- ROC Curve: Nos permite medir el desempeño general del modelo.

Resultados:
Al evaluar este modelo obtuvimos un desempeño general del modelo del AUC=86% menor que el obtenido con Naive Bayes con TF-IDF con AUC=93%. Además, comparando el balanced accuracy vemos que DistilBERT logró un 85% comparado con un 86% del primer modelo. Como nos interesa el desempeño al detectar reviews negativos, podemos notar que DistilBert logró un recall del 86% (capacidad de detectar las verdaderas reviews negativas) comparado con el 90% obtenido por Naive Bayes. Podemos concluir que Naive Bayes mostró un mejor desempeño en esta tarea de clasificación que DistilBERT.

Epoch	Training Loss	Validation Loss	Accuracy	Balanced Accuracy	Precision	Recall	F1
1	0.398500	0.369386	0.849000	0.849000	0.851107	0.846000	0.848546
2	0.321100	0.428723	0.814333	0.814333	0.906816	0.700667	0.790523
3	0.198400	0.510506	0.852333	0.852333	0.864734	0.835333	0.849780
4	0.209200	0.709780	0.844333	0.844333	0.888638	0.787333	0.834924
5	0.105100	0.718234	0.852333	0.852333	0.850697	0.854667	0.852677

```
Reporte de Clasificación:

              precision    recall  f1-score   support

           0       0.97      0.85      0.91     52797
           1       0.50      0.86      0.63      9027

    accuracy                           0.85     61824
   macro avg       0.74      0.86      0.77     61824
weighted avg       0.90      0.85      0.87     61824


Balanced Accuracy:

0.8552877081013338
```



3. VADER (Valence Aware Dictionary and sEntiment Reasoner)

Justificación:
VADER, diseñado para análisis de sentimientos en lenguaje coloquial (como redes sociales), se consideró adecuado debido al estilo informal de las reseñas. Además, este modelo proporciona un puntaje continuo de valence que permite identificar el grado de “positividad” o “negatividad” de un comentario, y a partir de un umbral permite clasificar entre comentarios positivos y negativos; de esta forma, no requiere un entrenamiento adicional con datasets específicos, y el fine-tunning se realizó con el valor de dicho umbral.

Metodología:

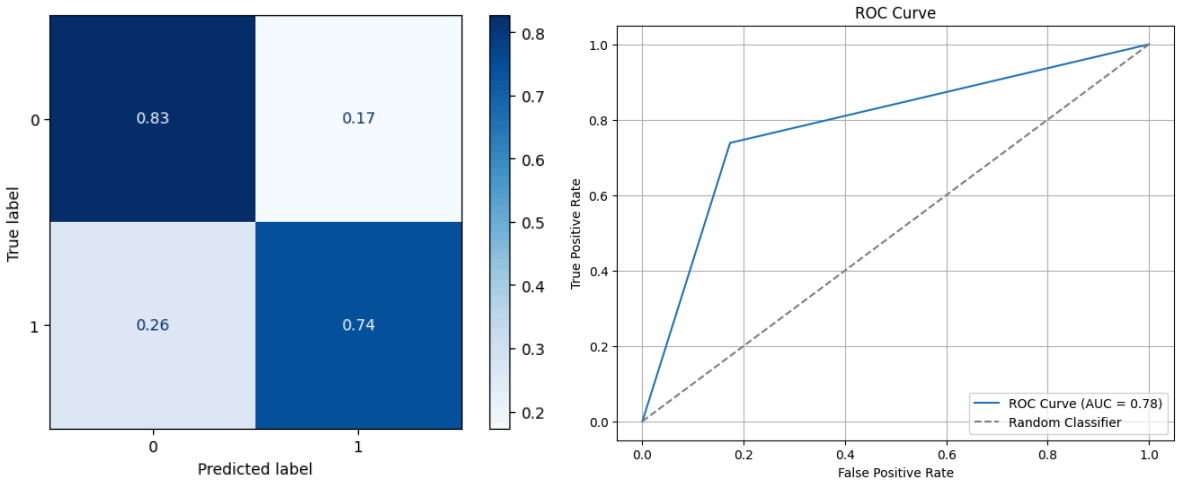
- Se probó tanto con datos limpios como sin limpiar.
- Se utilizó la variable compound (indicador de positividad o negatividad del texto).

Métricas de Evaluación:

- balanced_accuracy
- f1 score
- Sensibilidad
- Valor predictivo negativo

Resultados:
Aunque VADER está optimizado para textos informales, su desempeño fue inferior al modelo Naive Bayes con TF-IDF y a DistilBERT, especialmente en la identificación de reseñas negativas.

```
Mejor umbral: 0.7171717171717173
Mejor métrica: 0.7810692825239358
              precision    recall  f1-score   support

           0       0.95      0.83      0.88     52797
           1       0.42      0.74      0.54      9027

    accuracy                           0.81     61824
   macro avg       0.69      0.78      0.71     61824
weighted avg       0.87      0.81      0.83     61824
```



Modelos de Detección de Tópicos

Se exploraron tres enfoques principales para identificar temas en las reseñas negativas:

1. LDA (Latent Dirichlet Allocation)

- Métodos probados: Vectorización por conteo de palabras y TF-IDF.
- Resultados destacados:
    - Café: ['would', 'one', 'beans', 'good', 'flavor', 'cups', 'taste', 'like', 'cup', 'coffee'].
    - Entregas: ['bag', 'received', 'time', 'ordered', 'box', 'order', 'amazon', 'product'].

2. NMF (Non-negative Matrix Factorization)

- Enfoque: Modelo no probabilístico basado en descomposición matricial.
- Resultados destacados:
    - Té: ['drink', 'chai', 'black', 'weak', 'flavor', 'cup', 'bags', 'teas', 'green', 'tea'].

    - Chocolate: ['white', 'cocoa', 'milk', 'cookie', 'bar', 'dark', 'bars', 'hot', 'cookies', 'chocolate'].

3. BERTopic

- Descripción: Modelo LLM entrenado para detección de tópicos.
- Observación: Generó más de 100 grupos, aunque muchos no eran significativos.
- Resultados destacados:
    - China: ['dog', 'dogs', 'food', 'treats', 'china', 'made', 'product', 'treat', 'chicken', 'one'].

Análisis General:
Los temas recurrentes en las opiniones negativas incluyen:

1. Café: Problemas de sabor y calidad.
2. Té: Quejas sobre sabor y empaques.
3. Entregas: Retrasos y errores en pedidos.
4. Productos de China: Percepciones negativas sobre el origen.
5. Chocolate: Insatisfacción con el sabor o calidad.

Conclusión

El modelo Naive Bayes con TF-IDF demostró ser la mejor opción para la clasificación de reseñas negativas, mientras que LDA y NMF destacaron en la identificación de temas. Aunque herramientas como DistilBERT, VADER y BERTopic tienen aplicaciones prometedoras, no superaron las alternativas seleccionadas.

El análisis de tópicos revela áreas clave de mejora para la empresa, especialmente en productos específicos (café, té, chocolate) y procesos (entregas). Estas observaciones permiten priorizar las acciones para abordar las quejas más comunes y mejorar la experiencia del cliente.

Anexos

En esta sección de anexos, colocaremos explicaciones más detalladas que se pueden encontrar en la libreta de jupyter con el código.

EDA:

En el análisis exploratorio de los datos, nos dimos cuenta del gran desequilibrio en el conjunto de datos: la mayoría de las reseñas son positivas, mientras que solo el 14% son negativas. Además, limpiamos los datos eliminando registros duplicados y combinamos las columnas `Summary` y `Text` para tener una única columna, conservando toda la información.

En cuanto a la limpieza enfocada en PLN, eliminamos "stopwords", caracteres no alfanuméricos, espacios en blanco múltiples y tokenización. En particular todo lo que esté dentro de "<>", que son etiquetas de html que no nos interesan. Este es un procedimiento común en esta área. Por último, como se nos indicó en la descripción del proyecto, convertimos las reseñas con puntuaciones de 1 a 5 en una clasificación binaria, donde 1 representa las reseñas negativas y 0 las positivas.

No se aplicaron lematización ni "stemming" para evitar la pérdida de información relevante para nuestros modelos y también para reducir el tiempo de cómputo.

Naive Bayes:

El primer modelo que analizaremos es Naive Bayes con vectorización TF-IDF. Elegimos este modelo junto con esta técnica de vectorización debido al sorprendente rendimiento que demostró a lo largo del curso en diferentes tareas. Ahora veremos si en esta tarea de análisis de sentimiento logra mantener ese desempeño.

Para la obtención de los hiper-parámetros `max_features` y `class_prior`, utilizamos la experimentación, considerando tanto la métrica `balanced_accuracy` como el tiempo de cómputo. En este modelo, al igual que en los siguientes, estratificar los datos al momento de dividirlos, debido al notable desequilibrio en las clases. Escogimos la métrica `balanced_accuracy` como principal, ya que está diseñada específicamente para casos con clases desbalanceadas. Otra métrica

buena obtenida es la curva ROC, la cual nos dice que nuestro modelo es un buen clasificador.

Al finalizar nuestras pruebas, en el tiempo disponible, obtuvimos resultados satisfactorios, especialmente al considerar la rapidez y la simplicidad del modelo, características que también se observaron en otros ejercicios previos.

VADER:

Otra opción que exploramos fue VADER (Valence Aware Dictionary and sEntiment Reasoner), una herramienta de análisis de sentimientos basada en reglas y léxico. Esta herramienta fue entrenada para analizar sentimientos en mensajes de redes sociales, es decir, en lenguaje coloquial e informal (según su repositorio en GitHub, incluso puede reconocer emojis). Esta característica encajaba perfectamente con nuestro caso, ya que las reseñas contienen un lenguaje similar. Sin embargo, los resultados obtenidos no fueron satisfactorios y fueron inferiores a los de los modelos previos.

Realizamos pruebas utilizando tanto el conjunto de datos sin limpiar como el conjunto limpiado, con diferentes valores de "compound" (una variable que indica qué tan negativo o positivo es el sentimiento, donde -1 es extremadamente negativo y +1 es extremadamente positivo). También consideramos diversas métricas como `balanced_accuracy`, `f1 score`, `sensitivity` y `negative predictive value`. Las primeras métricas se enfocan en conjuntos de datos desbalanceados, mientras que las segundas priorizan identificar la mayor cantidad de verdaderos negativos, ya que este era un requisito de la tarea. Al final obtuvimos los mejores resultados con los datos sin limpiar y tomando en cuenta la métrica `balanced_accuracy`.

Aunque este modelo está diseñado específicamente para este tipo de datos, los resultados obtenidos con nuestro sencillo modelo de Naive Bayes con vectorización TF-IDF fueron superiores.

Análisis de tópicos:

Para la obtención de tópicos exploramos varios métodos, como LDA (Latent Dirichlet Allocation) con vectorización por conteo de palabras y TF-IDF, NMF (Non-negative Matrix Factorization), y finalmente un modelo LLM como BERTopic. La elección de estos modelos se basó en la

comparación de enfoques: LDA, visto en clases, es un modelo probabilístico, mientras que NMF es un modelo no probabilístico. Por último, BERTopic es un modelo LLM entrenado específicamente para la detección de tópicos.

Un comportamiento curioso de BERTopic es el número de grupos que genera cuando no se especifica previamente el número deseado; en este caso, produjo alrededor de 100 tópicos. Sin embargo, muchos de estos no ofrecían información significativa, al menos entre los que pudimos revisar.

Tras explorar los modelos, experimentar con ellos y compararlos, logramos identificar cinco productos o situaciones que generan la mayor cantidad de reseñas negativas:

1. Café
2. Té
3. Entregas
4. China
5. Chocolate

A continuación, presentamos algunos ejemplos de los tópicos generados por los diferentes modelos:

* LDA con Bag of Words:

    Tópico 0: ['would', 'one', 'beans', 'good', 'flavor', 'cups', 'taste', 'like', 'cup', 'coffee'] (Café)

    Tópico 2: ['sugar', 'really', 'would', 'one', 'cookies', 'flavor', 'good', 'chocolate', 'like', 'taste'] (Chocolate)

    Tópico 3: ['would', 'one', 'tastes', 'good', 'drink', 'water', 'flavor', 'like', 'taste', 'tea'] (Té)

    Tópico 5: ['bag', 'received', 'time', 'would', 'one', 'ordered', 'box', 'order', 'amazon', 'product'] (Entregas)

    Tópico 6: ['one', 'item', 'would', 'box', 'store', 'get', 'buy', 'amazon', 'product', 'price'] (Entregas)

* LDA con TF-IDF:

```
    Tópico 0: ['bitter', 'beans', 'good', 'weak', 'flavor', 'like',
'taste', 'cups', 'cup', 'coffee'] (Café)

    Tópico 1: ['one', 'ingredients', 'item', 'china', 'dog', 'buy',
'price', 'amazon', 'food', 'product'] (China)

    Tópico 3: ['sugar', 'one', 'good', 'tastes', 'drink', 'water',
'flavor', 'like', 'taste', 'tea'] (Té)

* NMF:

    Tópico 1: ['flavor', 'pods', 'beans', 'bitter', 'roast', 'instant',
'weak', 'cups', 'cup', 'coffee'] (Café)

    Tópico 2: ['drink', 'chai', 'black', 'weak', 'flavor', 'cup',
'bags',   'teas', 'green', 'tea'] (Té)

    Tópico 3: ['one', 'item', 'received', 'buy', 'ordered', 'order',
'price', 'box', 'amazon', 'product'] (Entregas)

    Tópico 4: ['chicken', 'made', 'cats', 'china', 'eat', 'treats',
'cat', 'dogs', 'food', 'dog'] (China)

    Tópico 5: ['white', 'cocoa', 'milk', 'cookie', 'bar', 'dark',
'bars', 'hot', 'cookies', 'chocolate'] (Chocolate)

* BERTopic:

    Tópico 0: ['product', 'amazon', 'box', 'price', 'order',
'received', 'item', 'ordered', 'would', 'one'] (Entregas)

    Tópico 1: ['taste', 'like', 'chocolate', 'flavor', 'product',
'good', 'one', 'would', 'cereal', 'eat'] (Chocolate)

    Tópico 3: ['coffee', 'cup', 'cups', 'like', 'taste', 'flavor',
'beans', 'good', 'roast', 'one'] (Café)

    Tópico 5: ['dog', 'dogs', 'food', 'treats', 'china', 'made',
'product', 'treat', 'chicken', 'one'] (China)


Un tópico que nos llamó la atención fue el de "China". Observamos que
muchas reseñas negativas estaban relacionadas con productos
```

provenientes de ese país. Algunas reseñas mencionan que les gustaban los productos hasta que descubren que eran de China. Otra observación interesante fue la asociación entre "perro" y "China". Inicialmente pensamos que podría deberse al estereotipo de que los chinos comen perros, pero en realidad, esto se debía a que muchos productos son alimentos para perros o porque los dueños mencionan que planeaban dárselos a sus perros debido a su origen de China.