SENTIMENT ANALYSIS

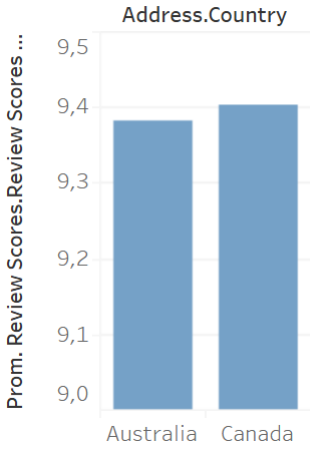
GORKA RUIZ, AIMAR NEGRO, IKER SALAZAR Y UNAI IGUARAN

Índice general

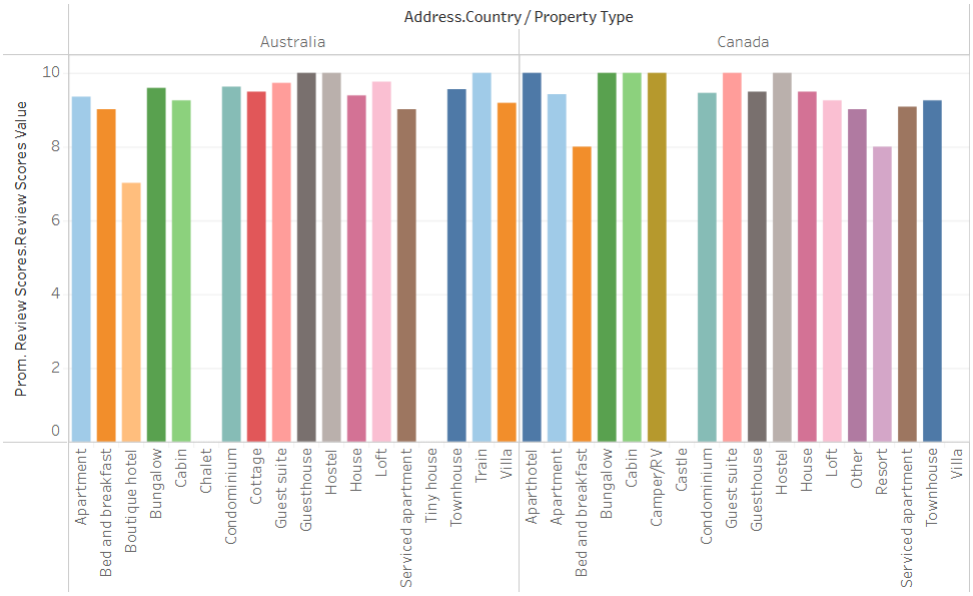
Índice de cuadros

1. Tableau: Análisis de los Datos iniciales

Nuestra zona es Canadá y la zona competidora es Australia. En los siguientes gráficos mostraremos las comparaciones de los datos de AirBBnB entre Canadá y Australia. Lo primero que hemos querido comprobar el promedio de las valoraciones de cada país.

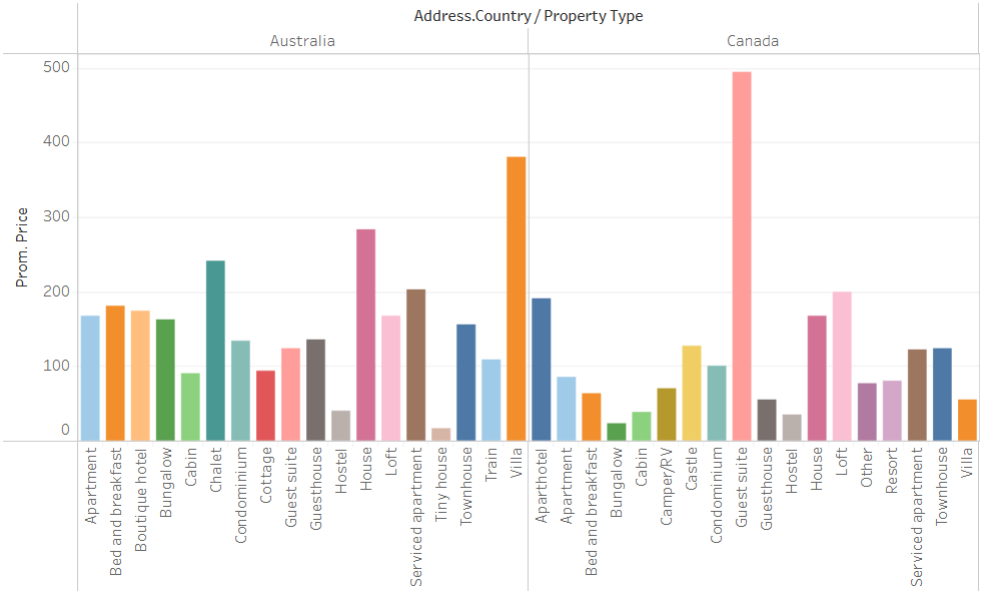


Tal y como se muestra en el gráfico, Canadá tiene por muy poco mejores valoraciones de AirBnB que en Australia. Canadá tiene aproximadamente 9,4 y Australia no terminar de llegar al 9,4. Continuando con las comparaciones, hemos decidido que era interesante mostrar qué tipos de propiedades eran más significativas en cada país.



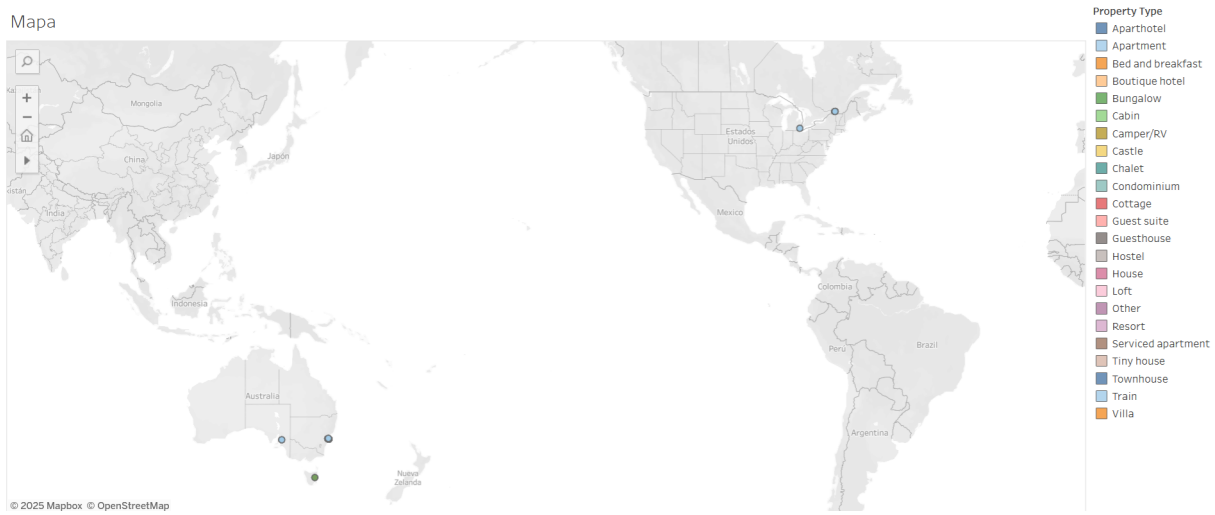
En Australia los alojamientos mejor valorados son las 'Guesthouse' y los 'Hostel' mientras que en Canadá son los 'Bungalow' y las 'Cabin'. Investigando sobre esto, hemos encontrado que se relaciona con el hecho de que en Australia, los turistas prefieren alojarse en las ciudades mientras que en Canadá se tira más a lo rural y querer estar junto a la naturaleza.

Luego, siguiendo con los apartamentos hemos querido comprobar si también tiene algo que ver con los precios de cada apartamento. Para ello hemos cogido el promedio de los precios de cada apartamento y los hemos comparado con el competidor.

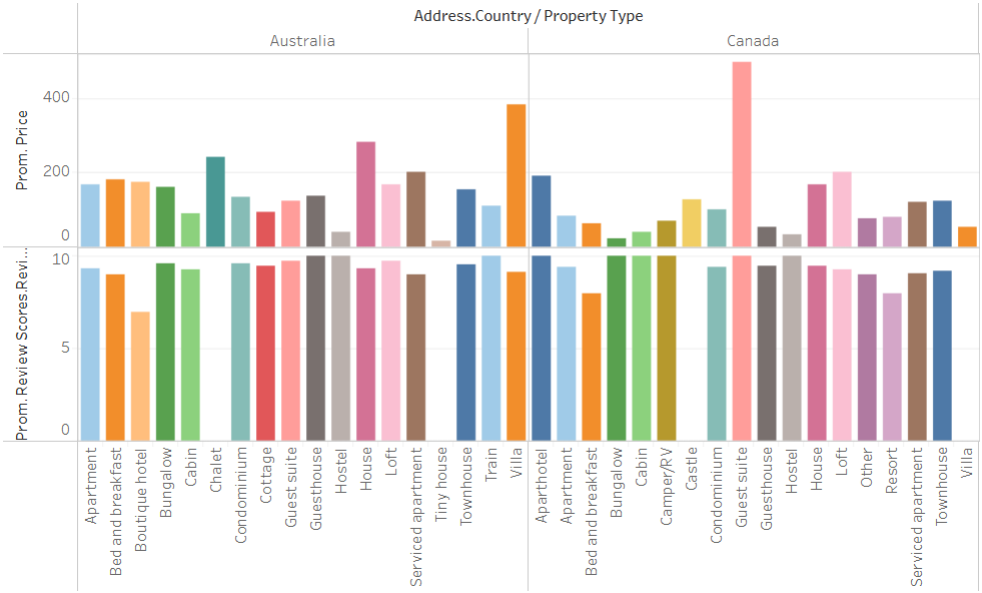


Entonces como se puede observar, justo el promedio de precios de los tipos de alojamientos mejor valorados tienen un precio más asequible que los demás.

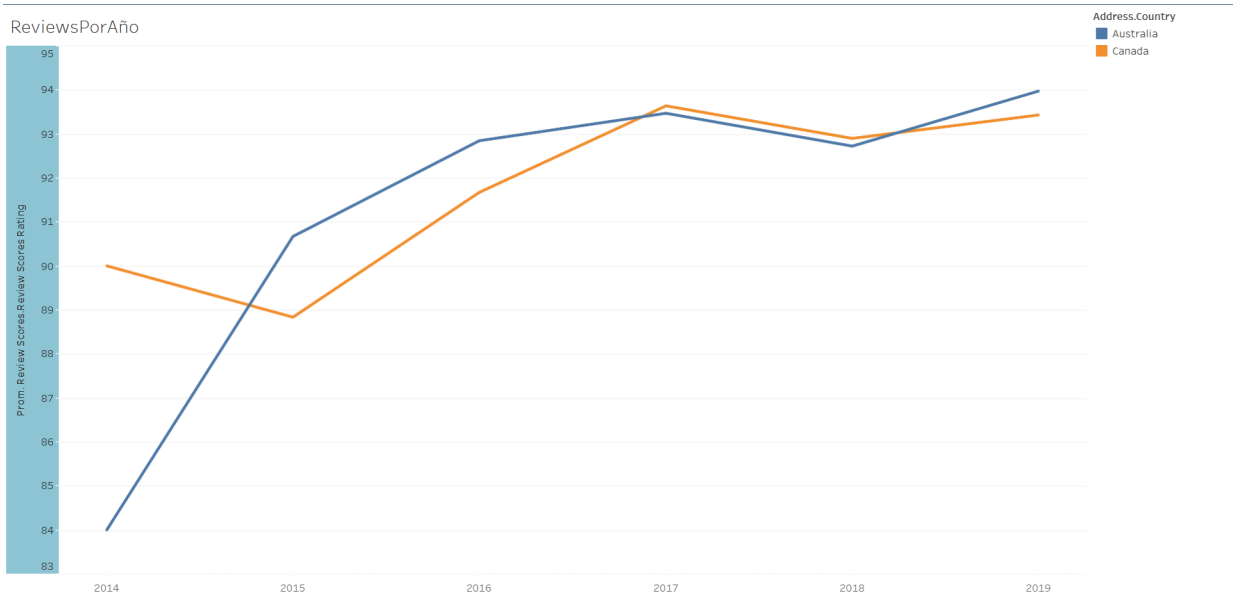
En el siguiente mapa se muestra por ciudades los tipos de apartamentos más alquilados en cada país. Se puede comprobar que los datos son muy céntricos.



Pasando al siguiente gráfico, podemos ver que generalmente cuanto más bajo es el precio mejores son las reviews salvo algunas excepciones en ambos paises.



En el siguiente gráfico mostraremos la valoración de los alquileres por año en los siguientes paises y veremos como hay un cambio significativo con el paso de los años.



Como podemos observar, en 2015 hay una diferencia grande entre Australia y Canadá ya que, mientras decrece en Canadá en Australia sigue creciendo. Al final se acaba equilibrando bastante las reseñas. Investigando sobre este caso, hemos encontrado que al parecer en Canadá en eso años había una regulación de alquileres a corto plazo muy estrictas al contrario que Australia, que tenía unas regulaciones mucho más laxas. Esto permitió a los anfitriones australianos operar con mayor flexibilidad, lo que podría haber contribuido a una mejor experiencia para los huéspedes y, por ende, a reseñas más favorables.

Tambien es interesante saber que la Comisión Australiana de Competencia y Consumo (ACCC) tomó medidas contra Airbnb en 2015 por prácticas de "precios escalonados", donde los costos adicionales no se mostraban claramente desde el principio. La fuente a esta información se puede comprobar en el siguiente enlace: referencia

2. Clasificación de sentimientos: Análisis, Preproceso y Experimentación

2.1. Graphical Abstract de la solución

2.2. Datos

2.2.1. División entre Train, Dev y Test de los datos para entrenar el modelo de predicción de ratings

Para entrenar y evaluar correctamente los modelos de clasificación que desarrollaremos, es fundamental realizar una adecuada partición de los datos disponibles. En esta sección se presenta la división de los datasets en tres subconjuntos: **entrenamiento (Train)**, **validación (Dev)** y **evaluación final (Test)**.

El objetivo de esta división es garantizar que los clasificadores puedan generalizar correctamente a datos no vistos. El conjunto de entrenamiento se utilizará para ajustar los modelos, el conjunto de validación servirá para comparar combinaciones de algoritmos e hiperparámetros, y el conjunto de test se reservará para la evaluación final del rendimiento de los modelos seleccionados.

Se trabajará con dos conjuntos de datos diferentes:

- **AirBnBReviews**: contiene reseñas etiquetadas como positivas o negativas, por lo que se utilizará para entrenar un **clasificador binario**.
- **tripAdvisorHotelReviews**: incluye valoraciones del 1 al 5, permitiendo el entrenamiento de un **clasificador multiclase**.

Además, se utilizará un conjunto central de datos, **AirBnB.csv**, en el que se evaluará la capacidad del clasificador para predecir correctamente la media de las valoraciones en un subconjunto específico. Es importante tener en cuenta que las escalas de ambos datasets pueden diferir, por lo que será necesario realizar un proceso de compatibilización entre ellas.

A continuación, se muestran las tablas con la división exacta de instancias para cada uno de los conjuntos, así como la distribución de clases en cada partición.

Conjunto De Datos	% de instancias	Num. de instancias
Train	60	204
Dev	20	68
Test Final	20	69

2.1. Cuadro: División Train, Dev y Test de los datos de AirBnBReviews

Conjunto De Datos	% de instancias	Num. de instancias
Train	70	14343
Dev	10	2049
Test Final	20	4099

2.2. Cuadro: División Train, Dev y Test de los datos de TripadvisorHotelReviews

Conjunto De Datos	% de instancias	Num. de instancias
Dev	80	8915
Test Final	20	2229

2.3. Cuadro: Dev y Test de los datos centrales del estudio que son los contenidos en AirBnB.csv

2.2.2. Distribución de las clases en cada conjunto

AirBnBReviews.csv

Conjunto De Datos	Clase Neg	Clase Pos.
Train	145	145
Dev	48	49
Test Final	49	48

2.4. Cuadro: Distribución Train, Dev y Test de AirBnBReviews

tripAdvisor_hotel_reviews.csv

Conjunto De Datos	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5
Train	995	1255	1529	4227	6337
Dev	142	179	218	604	906
Test Final	284	359	437	1208	1811

2.5. Cuadro: Distribución Train, Dev y Test de tripAdvisor

2.3. Preprocesamiento y Entrenamiento de Modelos

2.3.1. Objetivo

Se diferencian dos tipos de preprocesos, el primero para los datos que se van a utilizar para entrenar; el segundo para los datos centrales a predecir.

2.3.2. Preprocesamiento AirbnbReviews / TripAdvisorReviews

A cada conjunto de datos se le aplica un preprocesamiento independiente, que incluye los siguientes pasos:

1. Conversión del texto a minúsculas.
2. Eliminación de signos de puntuación, números y caracteres especiales.
3. Tokenización del texto.

4. Eliminación de stopwords.

5. Lematización para reducir las palabras a su forma base.

Una vez limpio el texto, se transforma en vectores numéricos utilizando técnicas de vectorización como **TF-IDF**.

2.3.3. Preprocesamiento Airbnb (Archivo Central)

A cada conjunto de datos se le aplica un preprocesamiento independiente, que incluye los siguientes pasos:

1. Se extraen las reviews de Canada.

2. Se exportan únicamente las reviews con su nota media asociada a otro csv.

3. Se traducen al inglés mediante el modelo generativo gemma2:2b, utilizando el siguiente prompt: *Eres un experto en traducción, la vida de muchas personas depende de tu trabajo. Traduce al inglés el siguiente texto manteniendo el significado original, solamente devolviéndome el texto traducido: Texto: {review} Traducción:.*

4. A la hora de predecir se le aplican los mismos preprocesos que a los datos utilizados para entrenar.

2.3.4. Primeros resultados de la tarea de clasificación

En esta sección se presentan los primeros resultados obtenidos en la tarea de clasificación de opiniones. El objetivo es evaluar el rendimiento de distintos algoritmos al enfrentarse a conjuntos de datos con características diferentes: uno de naturaleza binaria (`AirBnBReviews.csv`) y otro multiclase (`tripadvisor_hotel_reviews.csv`).

A través de los experimentos realizados, se busca comprobar la capacidad de generalización de los modelos entrenados, así como identificar los hiperparámetros que mejor se adaptan a cada escenario. Además, se explora de forma preliminar la aplicabilidad de estos modelos para predecir puntuaciones medias en un conjunto externo (`airbnb.csv`), lo que permitirá plantear una futura extensión del análisis.

En los apartados siguientes se recogen los resultados en forma de tablas para facilitar la comparación entre algoritmos, tanto en el dominio binario como multiclase.

Algoritmo	Hiperparámetros	Resultados (F-score)
Naive Bayes	Discretized alpha=0.5	0.97
KNN	k=5 p=1	0.81
RandomForest	n_estimators=10 max_depth=6 max_features=sqrt	0.89

2.6. Cuadro: Resultados sobre el Dev AirBnBReviews

Algoritmo	Hiperparámetros	Resultados (F-score)
Naive Bayes	Discretized `alpha=1.0`	0.426
KNN	`k=5, p=2`	0.362
Random Forest	`n_estimators=30`, `max_depth=50`, `max_features=sqrt`	0.382

2.7. Cuadro: Resultados sobre el Dev de los distintos algoritmos para Rating 1 (TripAdvisor)

2.3.5. Últimos resultados de la tarea de clasificación

En este apartado se presentan los útlimos resultados obtenidos en la tarea de clasificación. La estructura es la misma que en el apartado anterior.

Algoritmo	Hiperparámetros	Resultados (F-score)
Naive Bayes	Discretized alpha=0.5	0.97
KNN	k=2 p=1	0.89
RandomForest	n_estimators=10 max_depth=6 max_features=sqrt	0.89

2.8. Cuadro: Resultados sobre el Dev AirBnBReviews

Algoritmo	Hiperparámetros	Resultados (F-score)
Naive Bayes	MixedNB	0.421
KNN	`k=6, p=2`	0.382
Random Forest	`n_estimators=30`, `max_depth=50`, `max_features=sqrt`	0.39

2.9. Cuadro: Resultados sobre el Dev de los distintos algoritmos para Rating 1 (TripAdvisor)

Para hacer las predicciones en el archivo central Airbnb.csv se ha hecho un **ensamble** de los dos modelos entrenados. La fórmula utilizada para ello ha sido la siguiente:

$$s = x \cdot \text{Predicción}_{\text{TripAdvisor}} + \text{Predicción}_{\text{Airbnb}} + y$$

$$\hat{s} = \text{clip}(s,\, z,\, 10)$$

- x: Hiperparámetro de ponderación de la predicción del modelo entrenado con datos de `TripAdvisor`.
- y: Término de ajuste adicional (bias).
- z: Límite inferior aplicado al resultado final mediante la función `clip`.
- 10: Límite superior aplicado al resultado final mediante la función `clip`.
- $\hat{s}$: Resultado final tras aplicar el `clip` a s, limitado entre z y 10.

Algoritmo	Hiperparámetros	Rating medio predicho	Rating medio real	Desviacion
Ensamble	x=4.2 y=4.8 z=7	9.50	9.51	0.0001

2.10. Cuadro: Resultados sobre el dev y el test de airbnb final. Desviación=$(predicho - medio)^2$

2.3.6. Resultados de la aplicación de los modelos generativos

En esta sección se presentan los resultados obtenidos mediante la aplicación de modelos generativos a una tarea de predicción de valoraciones numéricas (del 1 al 10) a partir de reseñas textuales. A diferencia de enfoques tradicionales basados en clasificadores entrenados sobre conjuntos etiquetados como `AirBnBReviews.csv` o `tripadvisor_hotel_reviews.csv`, aquí se emplean modelos de lenguaje grande (LLM) para realizar predicciones directas sin necesidad de entrenamiento supervisado.

Se han evaluado diferentes estrategias de *prompting*: **zero-shot**, **one-shot** y **few-shot**. En este caso concreto, nos hemos centrado en la estrategia **zero-shot**, en la que se proporciona al modelo únicamente una descripción contextual junto con la reseña a valorar, sin ejemplos adicionales. Se han probado distintos *prompts* cuidadosamente diseñados para fomentar valoraciones positivas y minimizar respuestas ambiguas. El modelo empleado ha sido **Gemma2:2B**, ejecutado localmente mediante la plataforma `Ollama`.

El procedimiento ha consistido en:

- Evaluar diversos *prompts* sobre una muestra de desarrollo (*dev*) extraída del conjunto original.

- Medir el rendimiento de cada *prompt* en función de la **diferencia de medias** entre las puntuaciones reales y las predichas.
- Seleccionar el *prompt* con menor desviación para su posterior aplicación sobre el conjunto *test*.
- Analizar el comportamiento del modelo en términos de precisión media y desviación.

Prompts evaluados

A continuación se muestran los distintos *prompts* evaluados durante la experimentación. Cada uno de ellos fue diseñado para inducir al modelo a realizar una valoración numérica de una reseña entre 1 y 10, favoreciendo una puntuación positiva y evitando ambigüedad en la respuesta:

1. **Prompt 1**:

   > Eres un experto valorador de reseñas, la vida de mucha gente depende de tu respuesta. Valora esta reseña del 1 al 10 devolviendo solo el número, tira por lo alto, porfavor. Eres muy buena gente y te gusta poner muy buenas notas.
   > **Reseña:** {`review`}

2. **Prompt 2**:

   > Eres un crítico experto. Da una puntuación del 1 al 10 a esta opinión de usuario, tirando por lo alto. Eres muy buena gente y te gusta poner muy buenas notas. Responde únicamente con el número:
   > **Reseña:** {`review`}

3. **Prompt 3**:

   > You are an expert critic. Give this user review a score from 1 to 10, responding only with the number:
   > **Reseña:** {`review`}

Cada uno de estos *prompts* fue evaluado sobre el conjunto de desarrollo, y se seleccionó aquel que produjera una menor desviación media respecto a las valoraciones reales.

La Tabla 2.11 resume los resultados cuantitativos obtenidos sobre los conjuntos de desarrollo y prueba, mostrando los valores medios predichos y reales, así como la desviación asociada a cada configuración.

Modelo	Prompt	¿Rating medio predicho?	¿Rating medio real?	¿Desviacion?
Gemma2:2B	1	9.78	9.51	0.0729

2.11. Cuadro: Resultados sobre el dev y el test de airbnb final. Desviación=$(predicho - medio)^2$

2.4. Algoritmos, link a la documentación y nombre de los hiperparámetros empleados

2.4.1. Experimentación: Algoritmos empleados y Breve Descripción

Hemos empleado los siguientes algoritmos con los siguientes hiper-parámetros (no necesariamente tienen que ser estos sino los que empleeis).

- **Discretized Naive Bayes**:
  - Hiperparámetros: `alpha`
  - Link: Scikit-learn: MultinomialNB
- **Random Forest**:

- Hiperparámetros: `n_estimators=30`, `max_depth=50`, `max_features`
- Link: Scikit-learn: RandomForestClassifier

- **KNN (K-Nearest Neighbors)**:
  - Hiperparámetros: `k`, `p`
  - Link: Scikit-learn: KNeighborsClassifier

- **Modelos Generativos (Gemma 2B)**:
  - Hiperparámetros: No aplicables de forma directa como en modelos tradicionales; se puede ajustar la `temperature` y otros parámetros del motor Ollama.
  - Link: Ollama: Gemma 2B

2.5. Conclusiones y resultados

2.5.1. Discusión sobre el proceso de aprendizaje

Antes que nada creamos unos programas auxiliares para extraer las reviews de Canada y Australia. Además sacamos también las reviews de Canadá con su media asociada para poder calcular así mas fácil la media. Tras eso empezamos a pensar como abordar el experimento.

Se han realizado diferentes pruebas y pensamientos para llevar a cabo el aprendizaje.

De cara al preproceso decidimos decidimos primeramnete simplifcar el texto y luego aplicar **tf-idf**. A la hora de simplificar hubo signos de puntuación que decidimos no quitar, ya que pueden ser claros indicativos de sentimientos; como por ejemplo, :) o :(.

Una vez terminado el preproceso para ambos archivos empezamos a probar que algoritmos nos daban los mejores resultados.

Ambos modelos ofrecen las mejores métricas cuando se ha hecho uso del algoritmo de naive-bayes, por lo que ha sido el elegido. También hay que tener en cuenta que ese no ha sido el único criterio para el aprendizaje, ya que los resultados de estos modelos son ïntermedios", ya que lo que queríamos era obtener buenos resultados en el clasificador final.

Para ello nos dimos cuenta que en los datos de tripadvisor crear una columna por cada palabra era ineficaz, ya que había muchas y mermaba los datos. Para solucionar esto decidimos que para crear la columna debia repetirse un mínimo de veces. De esta manera, fuimos probando hasta que la media del clasificador final se aproximaba bastante a la media real. Lo curioso es que el modelo de tripadvisor para esta configuración daba peores resultados (ïntermedios"), pero nos ofrecía una media más aproximada.

También debemos tener en cuenta el **ensamble** que hemos utilizado, al principio pensamos en diferentes técnicas antes de llegar a esta.

Primero se ideó clasificar las reviews en dos grupos; las buenas, y las malas. Para saber si eran malas o buenas se utilizaría el modelo de aribnb (0 o 1), y luego se mapearian con los resultados del modelo de trip advisor. Es decir, si los resultados fueran 0 (Airbnb) y 2 (Tripadvisor) el resultado sería 2, si fuera 1 y 2 el resultado sería siete.

También se nos ocurrió buscar una operación que combinara los dos modelos para clasificar, pero no se nos ocurría alguna que ofreciera buenos resultados.

Finalmente, reciclando la segunda opción decidimos meterle hiperparámetros, para así obtener la fórmula que no encontrabamos. Con este método hemos conseguido acercanos a la media.

Para clasificar las reviews se extrajeron todas las reviews y se les asoció como nota el valor medio (score_value) de su grupo de reviews, no porque tengan esa nota, sino porque nos resultaba más sencillo ponerlo así para luego obtener la media, ya que no cambia (Este es el archivo que lee el clasificador final).

En github se encuentran todos los archivos utilizados para la realización del proyecto, incluidos los auxiliares: ProyectoSAD github

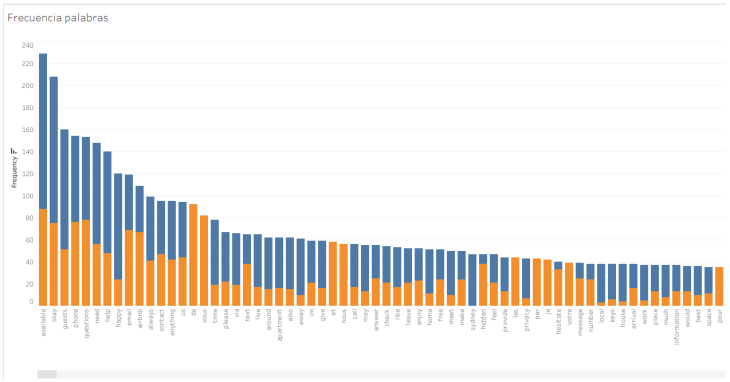
2.5.2. Conclusión sobre la tarea de clasificación

Teniendo en cuenta lo mencionado en el apartado anterior, esa fue la configuración que elegimos, ya que en el clasificador final es la que ofrece mejores resultados.

Marcamos como dato curioso que el obtener mejores resultados en algo no implica que sea bueno, ya que puede ofrecer peores resultados más tarde. Es importante definir que se quiere obtener y realizar los experimentos en base a ese objetivo, centrandose en obtener buenos resultados finales que intermedios.

2.5.3. Resultados Canadá VS Australia

En el siguiente gráfico mostramos la frecuencia de palabras por país, de color Azul Australia y de color Naraja Canadá.



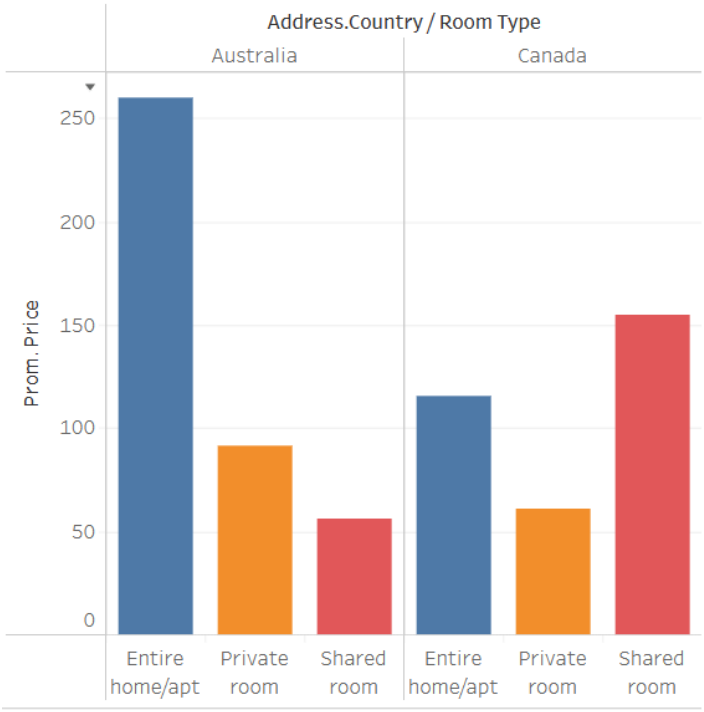
Dentro de las palabras más frecuentes hemos encontrado una que nos parece interesante y es "happy", que se repite más veces en Australia que en Canadá. Esto tampoco diferencia mucho, ya que no sabemos el contexto de esas palabras, pueden escribir "happyz que aún así sea mala la reseña. La palabra más repetida en ambos casos es .ªvailable", repitiendose en Australia más veces que en Canadá.

En general mirando todas las frecuencias, no hay ninguna diferencia a destacar o alguna palabra curiosa de comparar, ya que o bien hay frecuencia parecida entre palabras o directamente las palabras con más frecuencia no indican ninguna referencia a la estancia. Otro dato a tomar en cuenta es que las valoraciones eran ya bastante altas en los dos paises por lo que es difícil encontrar aspectos negativos en las palabras.

Mostrar gráficamente la distribución de los conceptos significativos.

2.5.4. Discusión sobre los descubrimientos realizados

La diferencia inicial que hemos encontrado sobre nuestro competidor es la preferencia de estancias. Estos se debe a que Canadá ofrece alquileres en zonas más rurales que en Australia que son más en las grandes ciudades. Si hablamos de precios, no hay una diferencia tan significativa pero si que cabe destacar, que en Canadá los precios son algo más bajos que en Australia en depende que tipo de alquiler.



Las valoraciones son muy parecidas aunque tiende a ganar Canadá. Lo único a tomar en cuenta es la diferencia con el paso de los años en cuanto al número de gente que optaba por AirBnB en Canadá y Australia. Como ya hemos explicado en el punto 1, esto se debía a las pocas regulaciones de alquileres del momento. Una vez fueron investigadas y aplicadas, los datos están mucho más parejos entre los dos.

Bibliografía

[1] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. Disponible en: `https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html`

[2] Scikit-learn. (2024). Random Forest Classifier. Disponible en: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`

[3] Scikit-learn. (2024). KNeighborsClassifier. Disponible en: `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html`

[4] Ollama. (2024). Gemma 2B Model Documentation. Disponible en: `https://ollama.com/library/gemma`

[5] Tableau Software. (2024). *Tableau Help - Official Documentation*. Disponible en: `https://help.tableau.com/current/guides/e-learning/en-us/tableau_overview.htm`