Stampone Rodrigo, Girod Joaquin, Ijjas Christian

Análisis de Sentimiento en Reseñas Online de Flybondi

Resumen

El servicio de las aerolíneas es un tema frecuentemente criticado por los usuarios, especialmente en situaciones de cambios de itinerarios o problemas operativos. Flybondi, una aerolínea low-cost en Argentina, ha sido objeto de diversas críticas en plataformas de reseñas online. Este trabajo tiene como objetivo aplicar técnicas de NLP a un conjunto de reseñas de clientes de Flybondi con el fin de identificar patrones y tendencias que permitan mejorar el servicio. A través del análisis de sentimiento, modelado de temas y visualización de los datos, buscaremos extraer información relevante sobre los aspectos más criticados por los clientes y ofrecer recomendaciones basadas en los resultados obtenidos.

Datos

Los datos a utilizar consisten en reseñas obtenidas de diversas plataformas de opiniones de clientes sobre Flybondi, tales como TripAdvisor, Google Reviews y foros especializados en viajes. Se hizo uso de técnicas de scraping para obtener los datos. El dataset incluirá el texto de las reseñas, la calificación numérica otorgada (cuando esté disponible), y cualquier metadato adicional que se pueda extraer.

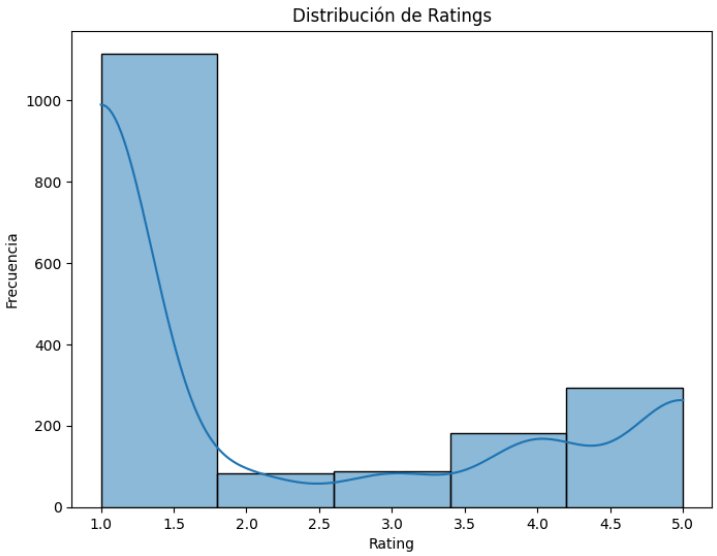
- **Fuente de los datos:** Reseñas públicas extraídas de plataformas como TripAdvisor, Google Reviews y Trustpilot.
- **Tamaño:** Aproximadamente 2000 reseñas.
- **Formato:** Los documentos estarán en formato CSV, con campos como “Texto de la reseña”, “Calificación” y otros metadatos.

La tecnología usada para extraer las reseñas fue un script de javascript muy simple que agarra elementos html y extrae su contenido. También se complementa con extensiones de Chrome que facilitan este proceso como la extensión “Instant Data Scraper” y la web https://outscraper.com.

Previo al análisis, se borraron registros que tuvieran reseñas nulas o duplicadas.

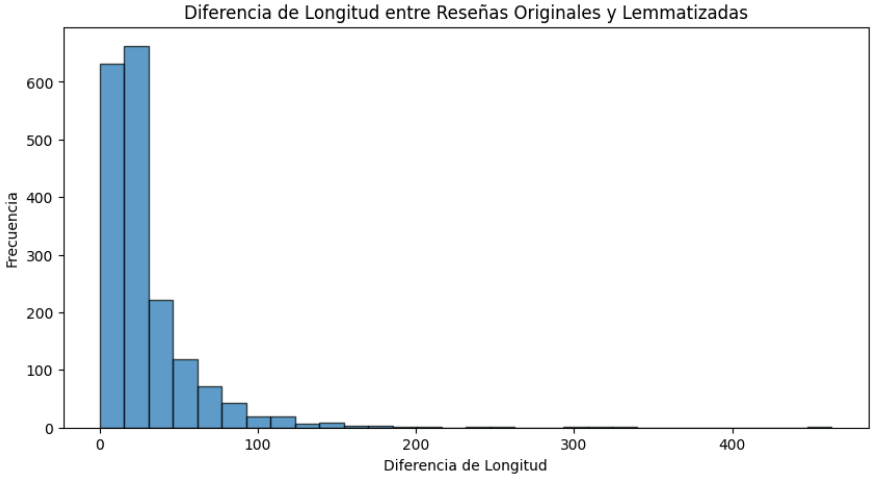
Análisis Exploratorio

Distribución de Ratings
La mayoría de las calificaciones (más del 50%) son de 1 estrella, con un menor porcentaje de críticas positivas. Esto refleja una polarización donde los usuarios suelen estar muy insatisfechos y nos da lugar a analizar el porqué del caso.



Lematizado vs Original

Se utilizó Stanza para lematizar las reseñas. Tras algunos desafíos iniciales en la instalación, se pudo completar exitosamente el proceso. Podemos ver que la mayoría de las reseñas presentan una diferencia pequeña en longitud tras el proceso de lematización. Hay algunas reseñas con diferencias en longitud que son más grandes, superando incluso los 400 caracteres. Esto podría deberse a reseñas que contenían muchas formas derivadas o variaciones de palabras, que fueron normalizadas al aplicar la lematización.



Wordcloud

Hicimos una vista de la distribución de palabras para ver cómo afectó el lematizador y se puede observar una clara mejora en cuanto al texto relevante. Inclusive así intentaremos limpiar mejor para destacar las palabras realmente relevantes ya que cosas como “flybondi” no nos interesan. Pero a simple vista se pueden destacar aspectos negativos en palabras como “desastre” y “nunca” entre otros.

Propuesta de Análisis

Buscamos proveer un análisis con mayores dimensiones de las que uno obtiene a simple vista con el promedio de ratings y el conjunto de opiniones positivas o negativas que suelen ofrecer los foros.

Este análisis podría ser de utilidad para Fly Bondi como punto de partida para dirigir planes de mejora, o podría ser utilizado por los foros para concentrar la alta cantidad de reviews en un texto más condensado y fácilmente navegable. Es importante tener en cuenta que un usuario que busca opiniones sobre una empresa rara vez extiende su exploración más allá de leer el promedio de reviews y algunas opiniones.

Además hipotetizamos que muchas de las quejas que se dirigen a Fly Bondi son injustificadas o basadas en eventos que están fuera de su control. Por ejemplo, las quejas por retraso suelen estar asociadas a problemas meteorológicos o sindicales, los cuales son de fuerza mayor para Flybondi, pero la gente opta por descargarse con la empresa de todas maneras.

Otro de nuestros objetivos es analizar cuáles son las falencias legítimas que tiene el servicio de Fly Bondi, para lograr este objetivo buscamos comenzar con un análisis de sentimiento pero abriendo la posibilidad a extender a un análisis de subjetividad al igual que un topic modeling sobre los temas recurrentes en las reviews.

En cuanto a técnicas, haremos uso de **Latent Dirichlet Allocation (LDA)** para identificar los temas recurrentes dentro de las reseñas. Esto nos permitirá descubrir cuáles son los aspectos del servicio de Flybondi que generan más discusiones, ya sea sobre puntualidad, comodidad, atención al cliente, etc.

Para facilitar la comprensión de los resultados, se harán visualizaciones como gráficos de barras, nubes de palabras y diagramas de dispersión que muestren claramente los temas más criticados y los sentimientos predominantes. Utilizaremos bibliotecas como Matplotlib y Seaborn para este propósito.

Experimentos

Experimento I

Preprocesamiento V2

En el TP1 realizamos el scraping de las páginas web que consideramos relevantes, y llevamos a cabo un proceso de lematización en el campo review_text. Sin embargo, tras un análisis más profundo, identificamos áreas donde el pre-procesamiento inicial de la información cruda podría mejorar significativamente.

Por esta razón, hemos desarrollado un pipeline de pre-procesamiento más robusto y completo, que incluye varias etapas adicionales para extraer y procesar mejor la información a partir de los campos originales. Esto nos permitirá obtener características más precisas y útiles para las siguientes fases del proyecto.

También cabe destacar que detectamos una fracción de las reviews que debido a una falla en el scrapeo habían limitado la cantidad de texto en las reseñas, por lo que procedimos a volver a scrapear correctamente esa sección.

Cabe mencionar que se decidió traducir las reseñas al idioma inglés, dado que la mayoría de las librerías estaban desarrolladas para este idioma. Si bien una traducción nunca va a llegar a ser del todo exacta, consideramos que la precisión obtenida con las herramientas para el idioma inglés compensa esta carencia.

Descripción del archivo final_combined_reviews.csv

El archivo final_combined_reviews.csv contiene los siguientes campos:

- **name**: Nombre del usuario que realizó la reseña.
- **experience**: Para las reseñas extraídas de Google, este campo muestra un título que resume la experiencia del usuario en función de la cantidad de reseñas y fotos que ha aportado al sitio. Ejemplo: "Local Guide . 16 reseñas . 60 fotos". En el caso de las reseñas extraídas de TripAdvisor y TrustPilot, este campo solo contiene la cantidad de opiniones del usuario.
- **rating**: La valoración numérica que el usuario otorgó en la reseña.
- **review_text**: El texto completo de la reseña escrita por el usuario.
- **likes**: El número de valoraciones positivas o "me gusta" que la reseña recibió por parte de otros usuarios.
- **review_title**: El título que el usuario asigna a la reseña.

Pipeline de Preprocesamiento

El código completo puede encontrarse en la notebook flybondi.ipynb, pero a continuación describimos los pasos principales del pipeline:

1. **Eliminación de filas duplicadas**: Se eliminan las filas que contienen reseñas duplicadas para evitar sesgos en el análisis.
2. **Normalización del campo rating**: Se unifica el formato de las valoraciones, que pueden aparecer como "5 estrellas", "1" o "1.0", para que todas contengan un único valor de tipo int.
3. **Normalización del campo likes**: El tratamiento es similar al del campo rating con la leve diferencia de que también se tiene que agregar el valor 0 a las filas vacías.
4. **Concatenación de texto**: Se combinan los campos review_text y review_title en un nuevo campo llamado review para centralizar toda la información textual de la reseña (algunas reseñas como las de tripadvisor contienen título además de la review).
5. **Generación del relevance_score**: Se crea una nueva métrica llamada relevance_score a partir de los valores del campo experience (incluyendo el título, la cantidad de reseñas y la cantidad de fotos) y el campo likes. A cada uno de estos factores se le asigna una ponderación parametrizable. Luego, la suma ponderada se normaliza con MinMaxScaling para obtener un valor final que refleje la relevancia de cada reseña.
6. **Remoción de emojis y puntuación**
7. **Identificación de Lenguaje y Traducción**: se utiliza un identificador de lenguaje para que luego el traductor pueda llevar todas las reviews al inglés.
8. **Remoción de Stopwords del Idioma Inglés**
9. **Lematización del campo review**: Se aplica lematización (del idioma inglés) al nuevo campo review, lo que permite reducir las palabras a su forma base, mejorando así la consistencia del análisis de texto.

Finalmente los campos restantes en el archivo de salida son: **name**, **rating**, **relevance_score**, **source_language**, **review**.

Experimento II

Análisis de sentimiento y subjetividad

Se decidió implementar un análisis de sentimiento y de subjetividad para extraer más información de las reviews y por ende conclusiones más acertadas.

Con el análisis de sentimiento, pudimos identificar si los comentarios de los pasajeros fueron positivos, negativos o neutrales, obteniendo los siguientes resultados:

```
sentiment
Negative    1061
Positive     713
Neutral       74
Name: count, dtype: int64
```

Fig. 1: Recuento de sentimientos.

En principio vemos una mayor parte de reseñas negativas, seguido por las positivas y por último las neutrales. El resultado de esto nos da también el Net Promoter Score (NPS) que es una

métrica muy útil para evaluar el mejoramiento de la retención de los usuarios y el porcentaje de gente que recomienda la aerolínea.

Los usuarios se dejan llevar por las emociones a la hora de opinar, y esta conducta suele aumentar aún más cuando se trata de servicios que los involucran directamente. Por esto un análisis de subjetividad puede ayudar a diferenciar las reseñas que son objetivas y permiten trazar planes de mejora para la empresa que se basen en los puntos inherentes al negocio.

Para la clasificación, se usarán números comprendidos entre 0 y 1, donde 0 es un comentario objetivo, y 1 un comentario subjetivo.

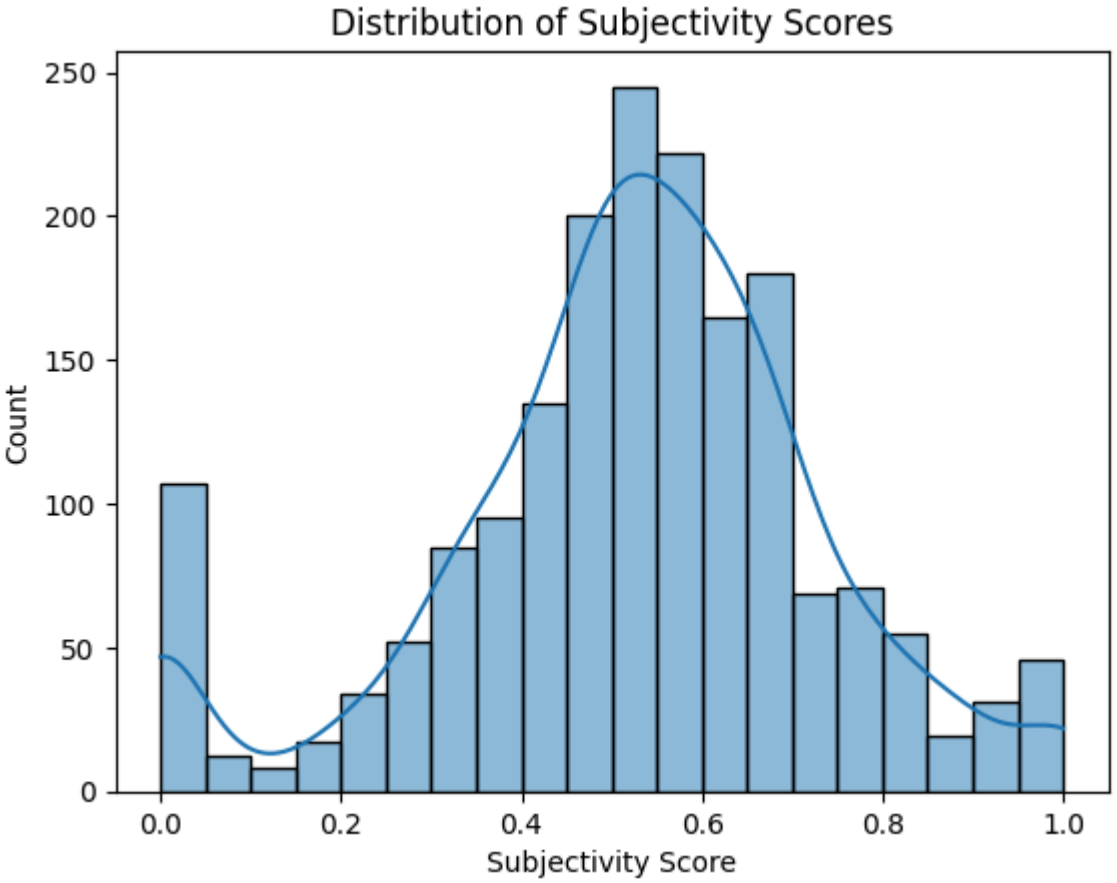


Fig. 2: Distribución de scores de subjetividad

Observando el gráfico, sigue una distribución similar a la normal, salvo que posee una importante acumulación de comentarios objetivos (0). Esto es clave, dado que los comentarios objetivos son lo que al analizarlos, podríamos llegar a encontrar puntos de mejora para la aerolínea. Sin embargo para poder apalancarse de esta conclusión es necesario hacer una lectura formal de las reviews que se taggearon como objetivas ya que algunas son tan carentes de información que producen información falsa. Como por ejemplo el caso de una review donde se comentó “Dd” y la traducción lo consideró del galés la traducción al inglés “goddess” y esto produce un objectivity score elevado.

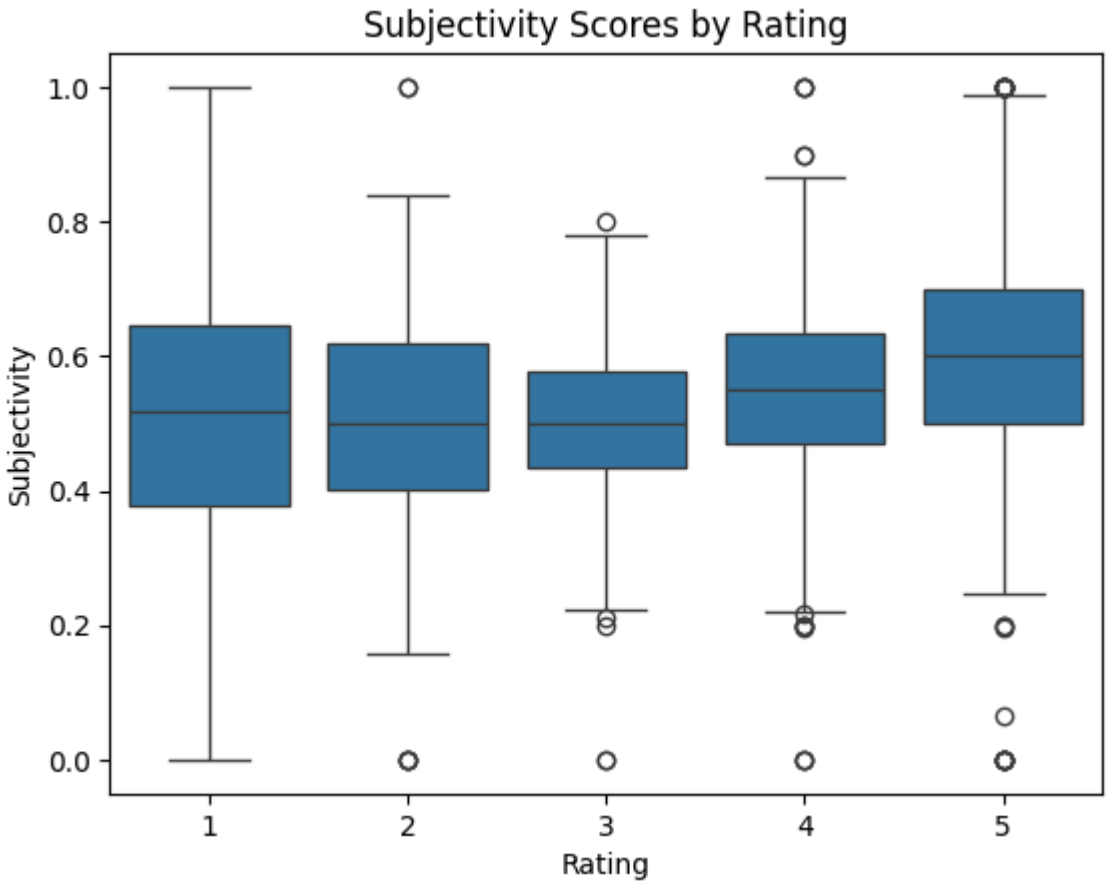


Fig. 3: Boxplot de subjetividad por rating

Al comparar la subjetividad contra el rating en la figura 3 se puede apreciar que los comentarios más subjetivos en promedio suelen ser los comentarios de mejor rating. También es interesante ver que las reseñas negativas son las que mayor variabilidad presentan, lo cual es congruente con la conducta esperada. A la hora de sacar conclusiones o hacer evaluaciones de desempeño tiene sentido descartar las reseñas más subjetivas tanto de las positivas como de las negativas, para poder basarse en evidencias empíricas.

Experimento III

Topic Modeling

A continuación, se decidió hacer uso de la técnica de Topic Modeling con LDA, marcando una diferenciación de los positivos y los negativos. Se marcó como positivas las reviews con rating mayor o igual a 4 y negativas los ratings menor o iguales a 2.

Cómo el objetivo del análisis es una mejora del modelo de negocios analizamos principalmente las reviews negativas. Con el Topic Modeling con LDA obtuvimos los tópicos y luego realizamos un WordCloud para ver visualmente las apariciones.

Fig. 4: WordCloud de los tópicos obtenidos mediante LDA

Estos tópicos quedan sujetos a interpretación pero a grandes rasgos se puede concluir fácilmente qué representan. Se probó diferenciar entre 5 tópicos en vez de 3 pero no había una separación clara.

Experimento IV

Análisis de sentimiento basado en aspectos

Decidimos utilizar un análisis de sentimiento basado en diferentes aspectos mencionados, para poder identificar áreas de mejora o puntos clave que la aerolínea realiza bien. Los aspectos definidos fueron: service, price, comfort, staff, food, flight y delay.

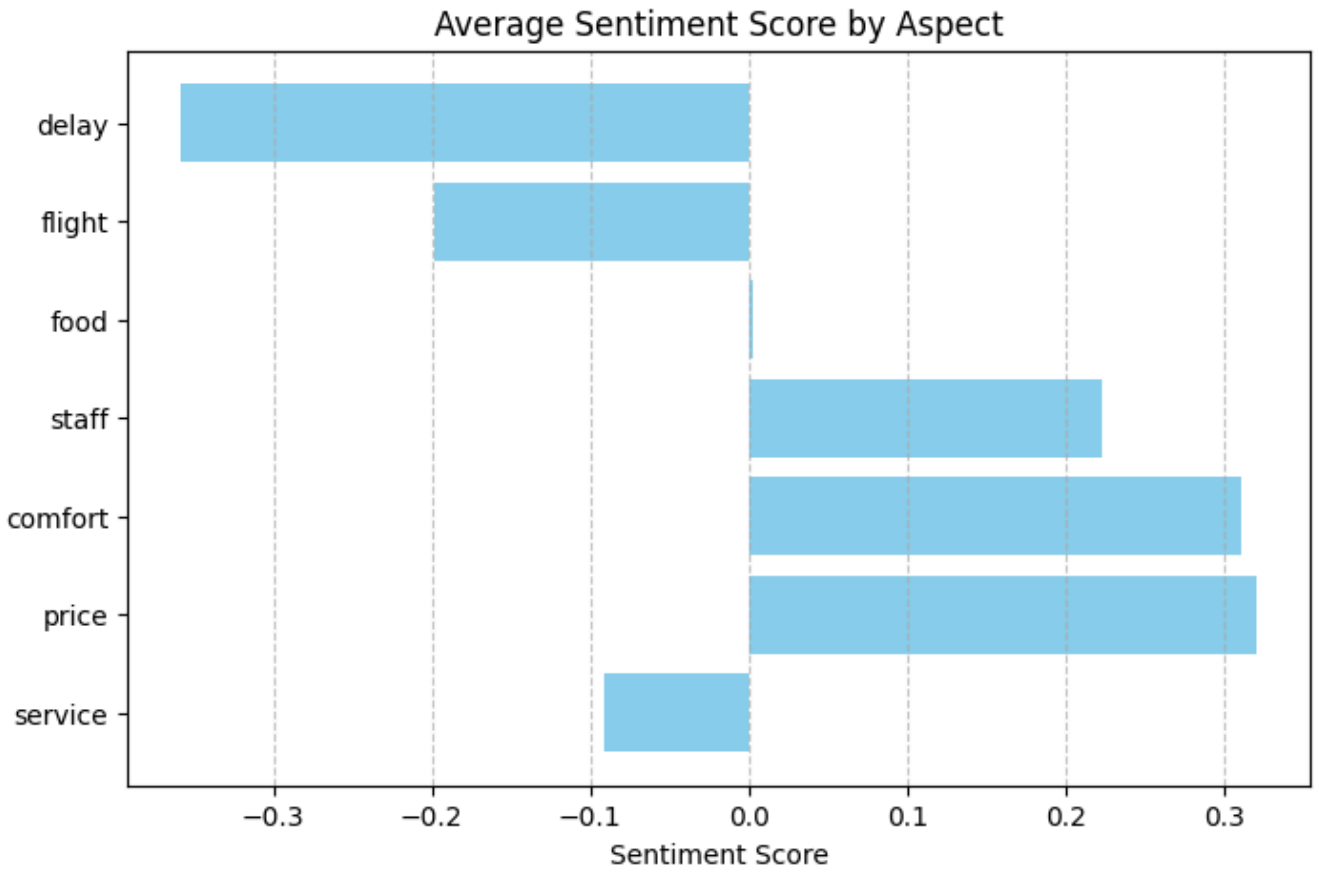


Fig. 5: Sentimiento promedio por aspecto

Nos encontramos con que el vuelo y el delay aparecen constantemente en comentarios negativos. Por otro lado, aspectos como el servicio, el precio o el staff suelen estar cumplidos de manera óptima. En base a esto, sería inteligente por parte de la aerolínea destinar recursos para solucionar de mejor manera los retrasos y la cancelación de vuelos. Si bien la mayor parte de las veces no depende de ellos, poder anticiparse y tomar algunas medidas mejoraría la satisfacción de los clientes con respecto a estos casos.

Tecnología usada

Todos los códigos están hechos en python y los datos se guardaron en csv.

- **Experimento I**
  - Dataframe manipulation: librería pandas
- **Experimento II**
  - Sentiment Analysis: VADER (Valence Aware Dictionary and Sentiment Reasoner), una herramienta de análisis de sentimientos que forma parte del paquete Natural Language Toolkit (nltk).
  - Subjectivity Analysis: TextBlob
- **Experimento III**
  - Topic modeling: LdaMulticore de Gensim
  - Word Cloud: librería WordCloud
- **Experimento IV**

- ABSA: nuevamente VADER

Próximos pasos

Consideramos que sería valioso avanzar con los siguientes temas

- Para el análisis de aspectos, hacer una lista curada basándonos en los tópicos del LDA y en la frecuencia de aparición de los lexemas, actualmente estamos usando una lista genérica de términos asociados.
- Experimentar con modelos multilingües y testear su desempeño contra los modelos en inglés, para decidir si vale la pena traducir un porcentaje más alto del dataset para poder utilizar modelos en inglés.
- Hacer un análisis predictivo para el sentimiento aprovechando el dataset etiquetado por medio de los ratings. Nuestra red se podría complementar a un modelo preexistente, y así posibilitar el análisis de texto proveniente de distintos canales menos estructurados (por ejemplo comentarios en las redes sociales)
- Generar el texto que resuma las opiniones generales de las reseñas.
- Hacer uso del relevance score para considerar más importantes las reviews de gente que tiene una mejor reputación online o experiencia en reseñas.
- Ponderar utilizando la objetividad de la review
- Mayor refinamiento de la data para los distintos modelos. Este parece ser el factor más importante que lleva a conclusiones buenas sobre los experimentos pues como se sabe: “**Garbage in, garbage out**”.