Project 2 report

Ruben Leon DIA3
Louis-Melchior Giraud DIA2 CDOF1

Github link :
https://github.com/Bluebloodfr/Project2_ML_NLP

Colab link :
https://colab.research.google.com/drive/1k592GB25_pbXOluiOktuSskM-VBqwC5X?usp=sharing

1. Data Collection and Loading

We began by gathering multiple Excel files, each containing user reviews related to insurance services. Using Python's os and pandas libraries, we iterated over the folder to read each .xlsx file and concatenated them into a single DataFrame. This unified dataset provided a comprehensive view of the reviews.

2. Data Cleaning and Preprocessing

- **Column Management**: We removed or renamed irrelevant columns such as author information, product type, and publication dates. We also handled missing values in the avis_en (English reviews) column by dropping rows with critical omissions. The note column (star rating) was assigned a default value of 0 where needed and cast as integer.
- **Splitting Train and Test**: We created two separate DataFrames:
    - A **training** set (with known star ratings).
    - A **test** set (where the star ratings were initially null and thus removed entirely).
- **Language and Labels**: The dataset included both French and English reviews. We set up configurations to process either French (avis_fr) or English (avis_en) by specifying the language of interest. We also defined two sets of labels (French and English) for topic classification.
- **GPU Check**: We included a check to see if a GPU was available (e.g., device = cuda) in order to speed up model inference during the next steps.
- **Highlighting Frequent Words**:
  We performed a frequency analysis of the most common words in the corpus using a Counter from the Python collections module. This helped us identify recurring terms such as *assurance*, *service*, and *prix* as top keywords.
- **N-gram Analysis**:
  We generated bigrams and trigrams to better understand common word pairs or triplets. For instance, *service client* and *tres satisfait* appeared frequently in French reviews, while *customer service* and *great experience* were prevalent in English texts.
- **Integration of a Spelling Correction Tool**:

To improve text quality, we utilized a dictionary-based approach with the pyspellchecker library (or TextBlob) to correct common misspellings in user reviews. This step ensured that our tokenization and subsequent NLP tasks (e.g., topic classification) were not disrupted by frequent typos.

3. Sentiment Analysis

We employed the **tabularisai/multilingual-sentiment-analysis** model because it handles both French and English inputs and outputs a sentiment score from 0 to 4. By adding 1 to the prediction, we aligned the final score with the star rating scale (1 to 5). The process involved:

1. **Tokenizing** each review.
2. **Predicting** a raw sentiment score between 0 and 4.
3. **Adjusting** by adding 1 to match our 1–5 rating scale.

A comparison of these predicted ratings to the ground truth star ratings showed an average absolute difference of approximately 0.79, suggesting a reasonable alignment between the model's sentiment predictions and user-assigned star ratings.

4. Topic Classification (Zero-Shot Approach)

We chose the **cross-encoder/nli-deberta-v3-base** zero-shot classification model to categorize reviews into predefined topics such as *Pricing*, *Coverage*, *Enrollment*, *Customer Service*, *Claims Processing*, *Cancellation*, and *Others*. Key steps included:

1. **Splitting Long Reviews**
   Reviews exceeding a specified word limit (approximately 500 words) were divided into chunks to ensure that important information was not truncated. We created an idmap so that each chunk still referenced its original review ID.
2. **Classification with a Confidence Threshold**
   We passed these chunks through the zero-shot model along with a list of labels (excluding the "Other" label at first). If the highest confidence score for any label fell below a certain threshold, we assigned the chunk to "Other."
3. **Merging Chunks**
   After classification, we recombined the chunked reviews by taking the average score of each label across all chunks of a single original review. The label with the highest average confidence became the assigned topic for that review.

This approach ensured that very lengthy reviews would not be improperly classified or truncated.

5. Summarization of Reviews

To derive concise representations of user feedback, we experimented with a summarization pipeline. We leveraged a transformer-based summarization model (for instance, **Falconsai/text_summarization**) and applied a chunk-based strategy:

1. **Chunk Creation**

    - We aggregated reviews into segments of manageable size (up to ~512 tokens or a chosen word count limit) to fit into the model's maximum input length.
    - If a single review exceeded this limit, we recursively split it until each chunk was below the threshold.
2. **Iterative Summarization**
    - Each chunk was summarized into a shorter text.
    - If there were multiple summarized chunks, they were merged and summarized again.
    - This iterative process continued until a single coherent summary was obtained for each selected group of reviews (e.g., all reviews from a particular insurer).

By applying this recursive technique, we created concise yet comprehensive summaries without losing essential details from longer texts.

6. Data Filtering

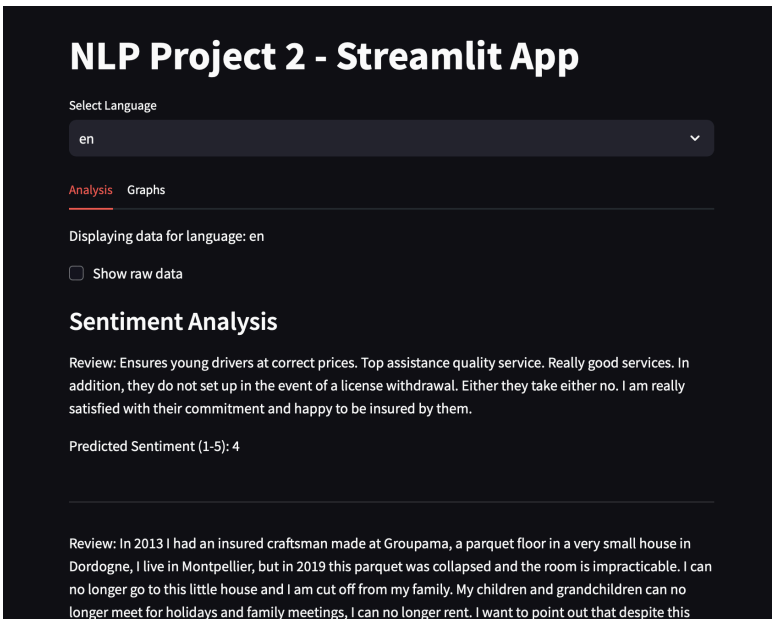
We developed a flexible filter function to select subsets of reviews based on specific conditions:

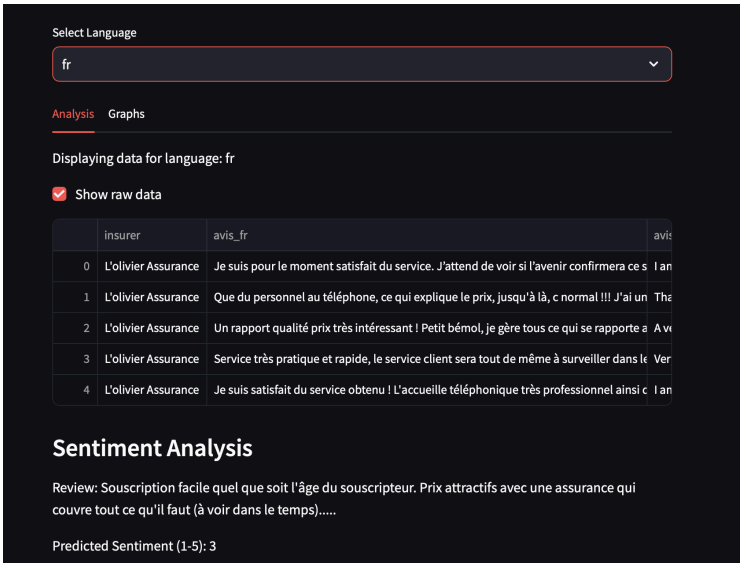
- **Insurers**: Filter by insurer name(s).
- **Labels**: Filter by topic (e.g., *Pricing* or *Customer Service*).
- **Star Ratings**: Filter by one or multiple star ratings (e.g., only 5-star reviews). This facilitated targeted analysis (e.g., retrieving all 5-star reviews about *Claims Processing* for a particular insurer).

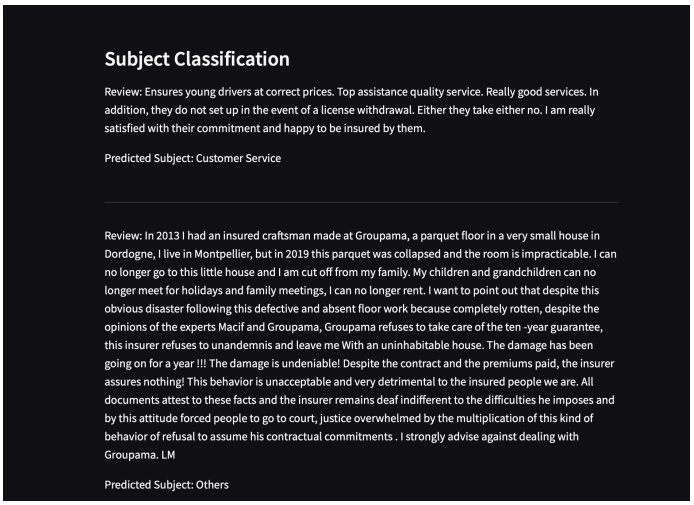
7. Streamlit Application

We developed a user-friendly Streamlit application that consolidates the outcomes of our data processing, sentiment analysis, topic classification, and visualization processes into an interactive interface. This application is divided into two main tabs—**Analysis** and **Graphs**—each serving distinct purposes. Below is an overview of the key features, along with indications where illustrative screenshots can be inserted.

7.1. Analysis Tab







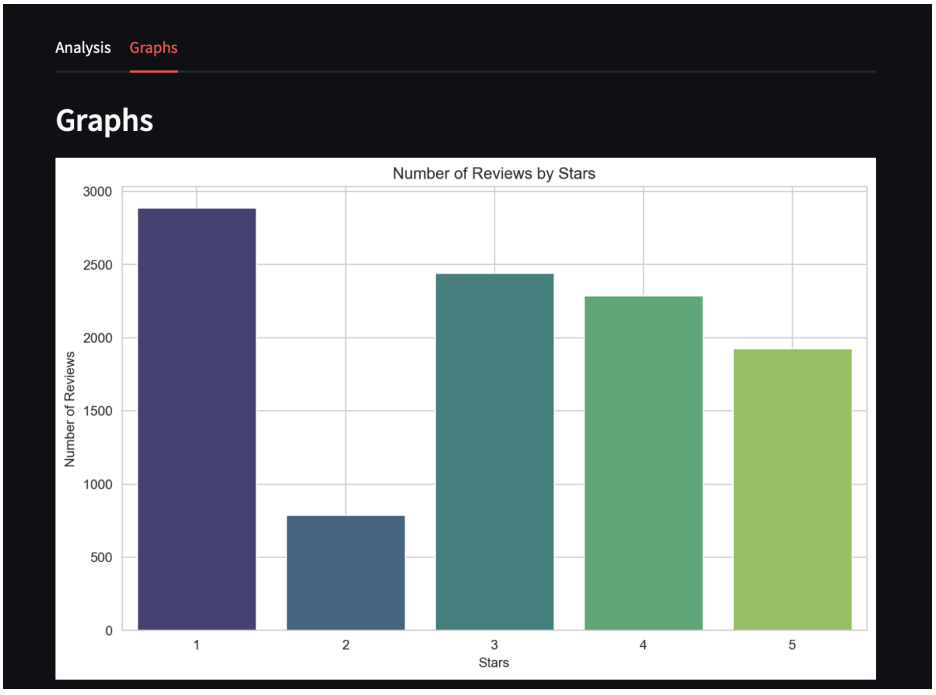
1. **Language Selection**
   At the top of the page, the user can choose the interface language (English or

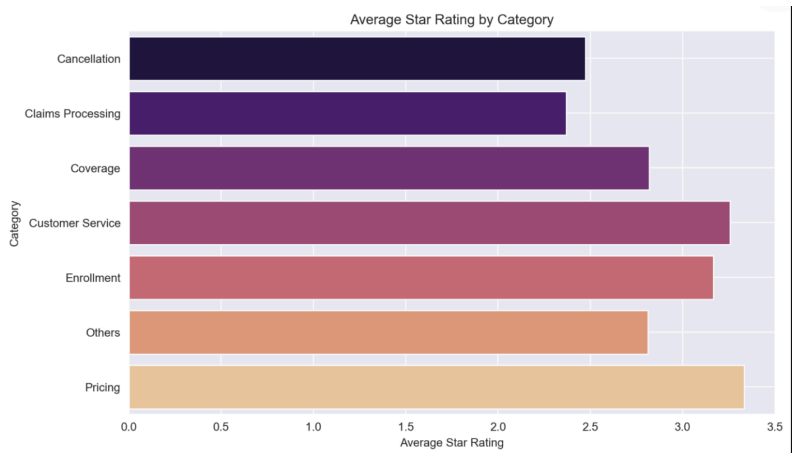
French) via a dropdown menu. Depending on the selection, the corresponding DataFrame (df_en or df_fr) is loaded in the background.

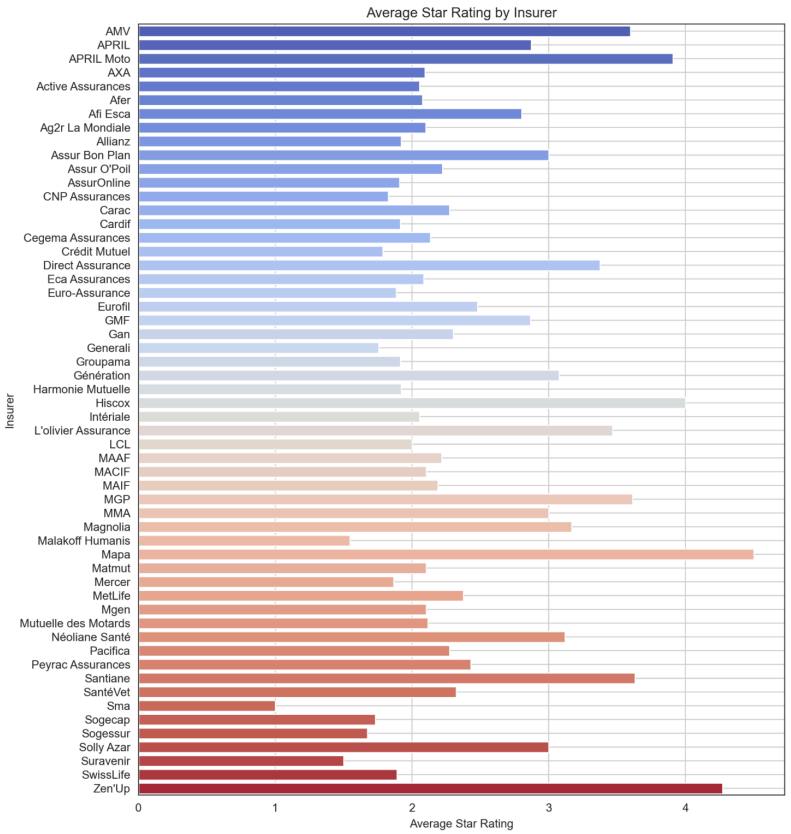
2. **Raw Data Display**
   By enabling a checkbox, users can preview the first five rows of the loaded dataset. This offers transparency into the contents of the DataFrame before any inference is performed.
3. **Sentiment Analysis**
   - **Sampling Reviews**: The application randomly selects five reviews from the chosen dataset.
   - **Prediction**: Each review is fed into our **multilingual sentiment analysis model**, which outputs a label between 0 and 4. We then adjust this score by adding 1 to align with a 1–5 star system.
   - **Results**: For each review, the predicted star rating (integer between 1 and 5) is displayed beneath the text.
4. **Subject Classification**
   - **Zero-Shot Model**: We employ a **Bart-based zero-shot classifier** to determine the most relevant topic among predefined categories such as *Coverage*, *Pricing*, *Customer Service*, etc.
   - **Thresholding**: If the highest confidence score for all topics is below a certain threshold, the model assigns the label *Others*.
   - **Output**: Users see the extracted review text along with the predicted subject classification.

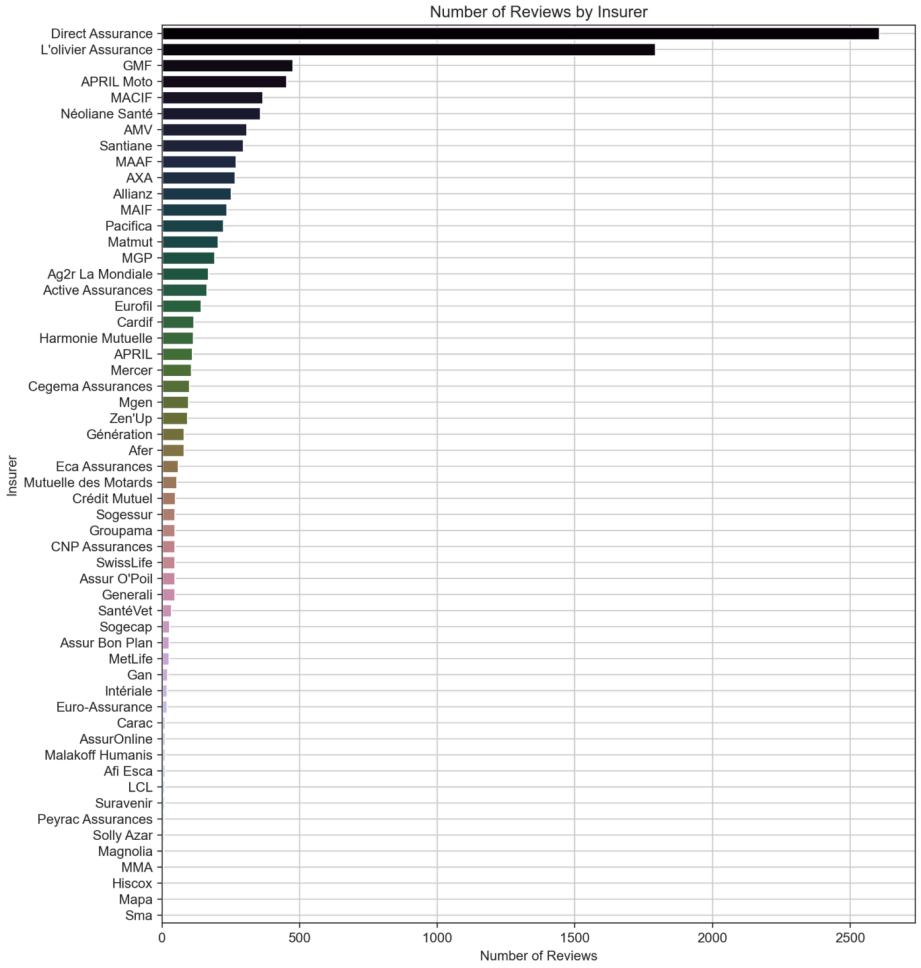
This tab enables a quick demonstration of how individual reviews are processed and classified, providing real-time insight into our model's performance.

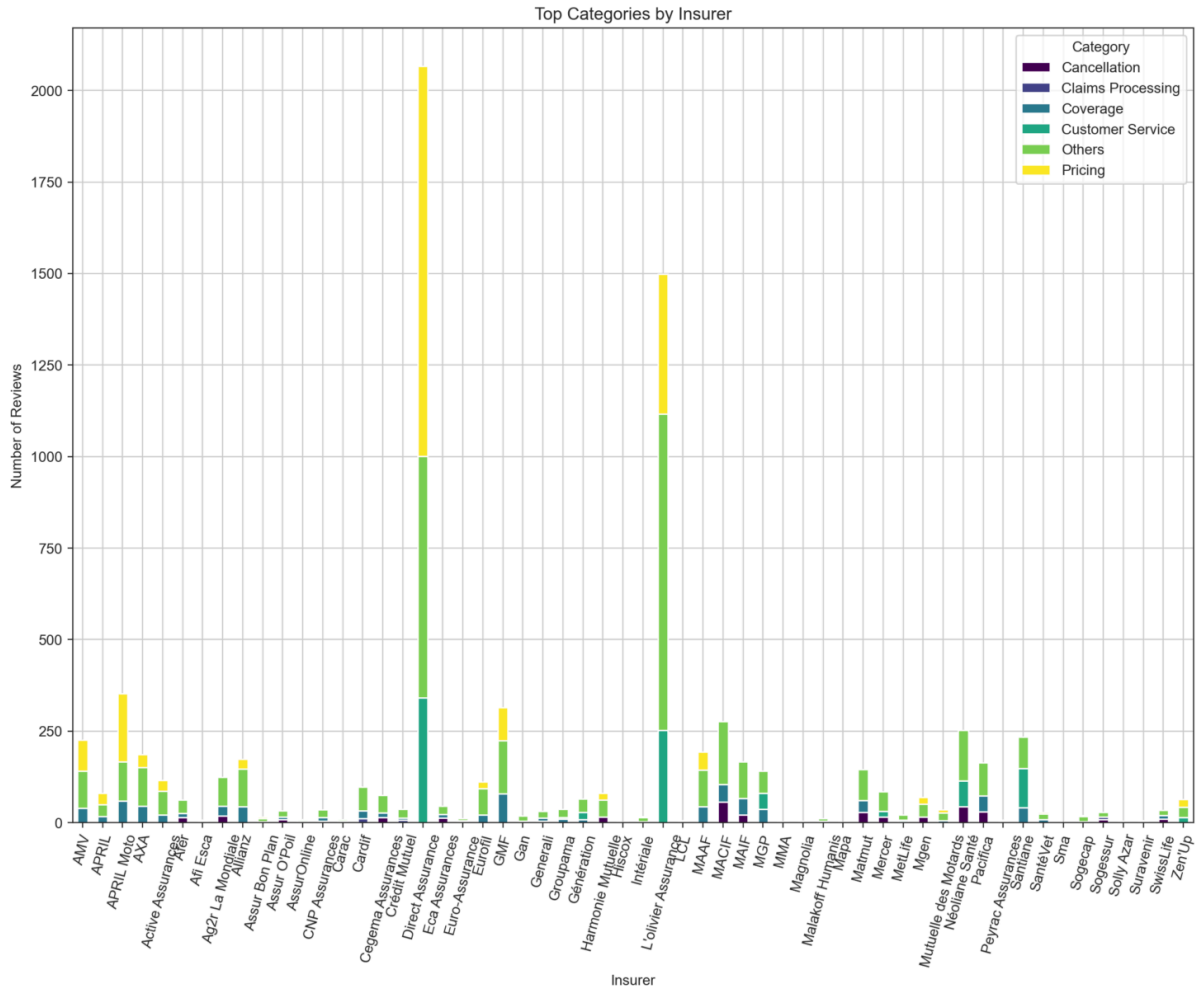
7.2. Graphs Tab




Average Star Rating by Category
Cancellation
Claims Processing
Coverage
Customer Service
Enrollment
Others
Pricing
Category
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Average Star Rating


Average Star Rating by Insurer
AMV
APRIL
APRIL Moto
AXA
Active Assurances
Afer
Afi Esca
Ag2r La Mondiale
Allianz
Assur Bon Plan
Assur O'Poil
AssurOnline
CNP Assurances
Carac
Cardif
Cegema Assurances
Crédit Mutuel
Direct Assurance
Eca Assurances
Euro-Assurance
Eurofil
GMF
Gan
Generali
Groupama
Génération
Harmonie Mutuelle
Hiscox
Intériale
L'olivier Assurance
LCL
MAAF
MACIF
MAIF
MGP
MMA
Magnolia
Malakoff Humanis
Mapa
Matmut
Mercer
MetLife
Mgen
Mutuelle des Motards
Néoliane Santé
Pacifica
Peyrac Assurances
Santiane
SantéVet
Sma
Sogecap
Sogessur
Solly Azar
Suravenir
SwissLife
Zen'Up
Insurer
0
1
2
3
4
Average Star Rating


Number of Reviews by Insurer
Direct Assurance
L'olivier Assurance
GMF
APRIL Moto
MACIF
Néoliane Santé
AMV
Santiane
MAAF
AXA
Allianz
MAIF
Pacifica
Matmut
MGP
Ag2r La Mondiale
Active Assurances
Eurofil
Cardif
Harmonie Mutuelle
APRIL
Mercer
Cegema Assurances
Mgen
Zen'Up
Génération
Afer
Eca Assurances
Mutuelle des Motards
Crédit Mutuel
Sogessur
Groupama
CNP Assurances
SwissLife
Assur O'Poil
Generali
SantéVet
Sogecap
Assur Bon Plan
MetLife
Gan
Intériale
Euro-Assurance
Carac
AssurOnline
Malakoff Humanis
Afi Esca
LCL
Suravenir
Peyrac Assurances
Solly Azar
Magnolia
MMA
Hiscox
Mapa
Sma
Insurer
0
500
1000
1500
2000
2500
Number of Reviews



The Graphs Tab offers a comprehensive **data visualization module**, enabling stakeholders to explore aggregated metrics and trends in the dataset. The following visualizations are available:

1. **Number of Reviews by Stars**:
    - A bar chart showing the distribution of reviews by their star ratings.
    - Insights:
        - The largest proportion of reviews have a 1-star rating, indicating a prevalence of dissatisfaction among users.
2. **Average Star Rating by Category**:
    - Displays average star ratings grouped by predicted topics such as *Pricing*, *Customer Service*, and *Cancellation*.
    - Insights:
        - Categories like *Pricing* show relatively higher satisfaction, while *Cancellation* scores the lowest.
3. **Average Star Rating by Insurer**:
    - A bar chart displaying the average star ratings per insurer.
    - Insights:
        - Insurers such as **"Zen'Up"** and **"APRIL"** are among the top-performing companies based on average ratings.

        - Companies like **Direct Assurance** and **L'Olivier Assurance** have a high number of reviews but relatively lower ratings.
4. **Number of Reviews by Insurer**:
    - Highlights insurers with the highest review volumes.
    - Insights:
        - **Direct Assurance** and **L'Olivier Assurance** dominate in review count, indicating their market presence or user engagement levels.
5. **Top Categories by Insurer**:
    - A stacked bar chart showing the frequency of predicted topics (e.g., *Pricing*, *Cancellation*) per insurer.
    - Insights:
        - *Customer Service* is a dominant topic for most insurers, but specific insurers like **Direct Assurance** have a significant focus on *Pricing*.

These visualizations enable stakeholders to quickly identify patterns and trends—such as which insurers receive the most reviews, how user sentiment aligns with specific insurers or topics, and the most common categories of user feedback.

By combining **real-time sentiment and topic predictions** (Analysis Tab) with **aggregated statistical overviews and visualizations** (Graphs Tab), this Streamlit application offers a comprehensive environment for exploring and understanding large sets of insurance-related reviews.

8. Observations and Insights

1. **Sentiment Analysis**:
    - The majority of reviews have polarizing sentiments, with many users rating either extremely low (1) or extremely high (5).
    - The sentiment analysis effectively captures user emotions, but some reviews classified as neutral (3) may benefit from further fine-tuning.
2. **Topic Classification**:
    - The zero-shot classification model successfully identifies key topics. However, the high frequency of "Others" suggests the need for more granular or contextual category definitions.
3. **Graphical Insights**:
    - Visualizations reveal trends in user satisfaction and allow insurers to pinpoint areas requiring improvement (e.g., dissatisfaction with *Cancellation* processes).
    - Companies with higher review volumes tend to have more diverse feedback, potentially influencing their average ratings.

In summary, our project tackled data consolidation, cleaning, sentiment analysis, zero-shot topic classification, and a chunk-based summarization pipeline. We integrated these components into a Streamlit application to provide an interactive platform that can predict star ratings, classify topics, and produce summaries of user feedback. This comprehensive approach empowers users to gain deep insights from large volumes of insurance-related reviews in a clear, intuitive manner.