Dataset 2

We conducted **Sentiment Analysis** to understand customer feedback trends (**Positive, Neutral, Negative**). Most feedback is **neutral**, followed by **negative**, with fewer **positive** responses.

We performed **Exploratory Data Analysis (EDA)** to understand the structure of the data before applying machine learning and NLP techniques.

Sentiment Analysis Approach

- Used **Google's Gemini Pro** with a predefined **JSON schema** to structure responses.
- Provided labeled conversation examples to guide the model in classifying sentiment.
- The model generates a structured **JSON output** with:
  - **Thoughts** (reasoning for classification)
  - **Sentiment** (one of: neutral, positive, negative, frustrated)
- High accuracy in frustrated cases indicates strong alignment with refund-related queries.

Dataset 1

The dataset is from Kaggle.

- Performed initial inspection steps and cleaned the data: Handled missing values and dropped irrelevant columns.
- Examined categorical features such as ticket type, priority, and language distribution. The dataset contains **four types of tickets**: incidents, requests, problems, and changes.

We analyzed relationships between **tags and ticket priority** and mapped tags to business types.

We also analyzed **language-specific patterns**:

- **Tag_1 Distribution**: Top 10 most used tags.
- **Queue Distribution**: Ticket volumes across different queues.
- **Missing Values Heatmap**: Identified sparsity in certain columns.

Response Automation

- Preprocessed text using **TF-IDF vectorization**, which helps extract important terms while reducing the impact of frequently occurring words.

- Used **PCA** to reduce high-dimensional **TF-IDF vectors to 3D** for visualization. This helps in identifying patterns in a lower-dimensional space.
- Employed **SentenceTransformer for embeddings**, which captures **semantic meaning** beyond simple word frequency-based methods.
- Used **NER for topic identification**, then grouped similar tickets into clusters using **K-means clustering** for efficient ticket categorization. Named clusters for better issue classification.
- Calculated **Percentage Coverage of Each Cluster**
  Top Issues within Each Cluster:
- **Printer Cluster:** Frequent issues related to **paper jams, wireless setup, and printer settings updates** (~3.64% for each).
- **Jira Cluster: Server configuration changes, Gmail syncing issues, new project setup** (~1.01% each).
- **AWS Cluster: Billing discrepancies, performance issues, deployment failures** (~1.50% each).
- **Dell XPS Cluster: Excel crashes, battery issues, screen problems** (~2.17% each).
- **Cisco Router Cluster: Critical network issues with ISR4331 routers** (~2.74% each).

AWS and Cisco Router clusters contain urgent issues, suggesting a need for faster response times.

Printer and Dell XPS clusters have recurring hardware and setup issues, indicating user education or better documentation might help.

Jira issues involve configuration and syncing problems, which might need better onboarding guides or support automation. So this data analysis can help us to prioritize support ticket resolution