**EDA Insights & Methodologies**

**Methodologies Used in the Project**

1. **Data Ingestion & Cleaning**
   * **Object-Oriented Pipeline:** We implemented classes such as DataLoader, DataCleaner, and DataTransformer to load JSON data into a Pandas DataFrame, drop duplicates, fill missing values, and handle unhashable fields by selecting only relevant columns for duplicate checks.
   * **Categorical Encoding:** We converted the config field into categorical codes to make them suitable for numerical analysis.
   * **Text Preprocessing:** Each message was tokenized, lowercased, had non-alphabetic characters removed, and was lemmatized. Stopwords were removed using NLTK’s English stopword list.
   * **Message Length Computation:** For each message, we calculated its word count and stored it in the message\_length field.
2. **Exploratory Data Analysis**
   * **Message Counts, Sentiment Distribution, Message Length Distribution:** We aggregated the data by agent to visualize how often each agent spoke, how they expressed sentiments, and how long their messages were.
   * **Word Cloud Generation:** We used the WordCloud library to visualize frequently used terms across all processed messages.
   * **Correlation Analysis:** We introduced a numeric feature (avg\_message\_length) to investigate relationships with other numerical columns, such as config\_encoded.
3. **API & Visualization**
   * **FastAPI Integration:** A REST API was built to serve the data pipeline functionalities, offering endpoints for real-time text preprocessing and transcript insights.
   * **Matplotlib & Seaborn Plots:** Graphs were generated in Python, saved as base64-encoded images, and embedded into HTML templates for a user-friendly EDA dashboard.
4. **Deployment Readiness**
   * **Dockerized Environment:** A Dockerfile allows the entire project to be containerized, ensuring consistent deployment across environments.
   * **Asynchronous Endpoints:** FastAPI’s asynchronous capabilities allow scalable handling of incoming requests, beneficial for future expansions of the dataset.

**EDA Insights**

**1. Message Counts per Agent**

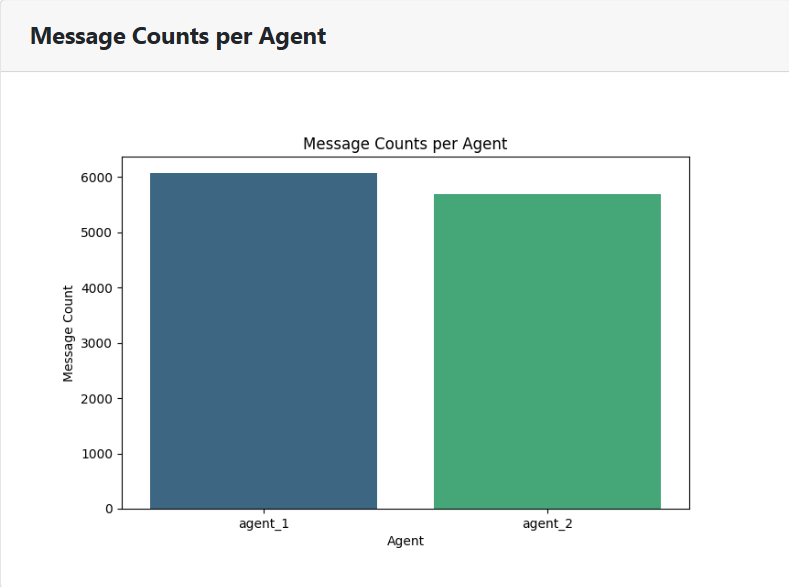
**Graph Description:**  
This bar chart shows how many messages each agent sent during the conversations.

**Insights:**

* **Comparative Participation:** If one agent has a higher bar, it indicates they contributed more messages overall. For example, if Agent 1’s bar is taller than Agent 2’s, Agent 1 might be driving the conversation more frequently.
* **Conversation Flow:** A more balanced count between agents suggests a two‐way discussion, while a skewed count implies one agent was more dominant or talkative.

**Potential Impact:**

* A higher message count for one agent could reflect that agent’s role (e.g., a support agent offering more explanations).
* Understanding who leads the discussion can inform training or conversation flow improvements.



**2. Sentiment Distribution per Agent**

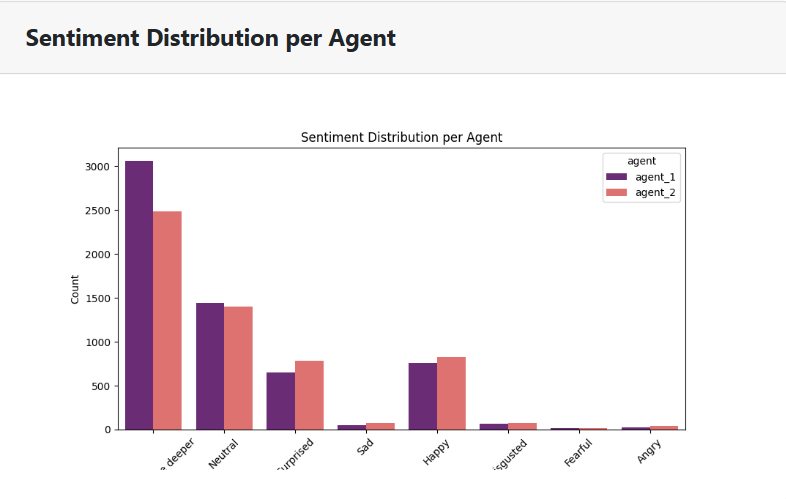
**Graph Description:**  
This histogram (or bar chart) shows how often each sentiment category appears, grouped by agent.

**Insights:**

* **Sentiment Variety:** If certain sentiments (e.g., “Happy,” “Sad,” “Neutral,” etc.) are more common for a particular agent, it can suggest that agent’s emotional style or the conversation context they handle.
* **Comparison Across Agents:** If Agent 1 has mostly “Curious” or “Happy” sentiments, while Agent 2 has “Neutral” or “Sad,” it might indicate differences in perspective, roles, or content they discuss.
* **Potential Bias:** If negative sentiments dominate, it could signal frustration or dissatisfaction within the conversations.

**Potential Impact:**

* Identifying high levels of negative sentiments might prompt deeper investigation into conversation topics or agent training.
* A more positive distribution could indicate effective communication or a less confrontational topic.



**3. Message Length Distribution per Agent**

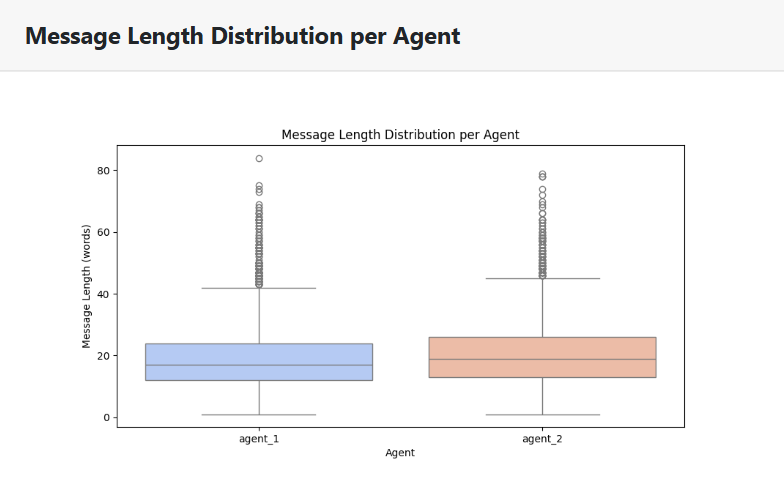
**Graph Description:**  
This box plot visualizes the spread of message lengths (in words) for each agent.

**Insights:**

* **Median & Spread:** The box indicates the median and interquartile range. If Agent 1’s box is narrower, that agent tends to use more concise messages. A wider box suggests more variability.
* **Outliers:** Points lying outside the whiskers are outliers—especially lengthy or short messages that differ from typical conversation.
* **Comparison:** If one agent consistently has shorter messages, it may reflect the role (e.g., short clarifications or questions). Longer messages might indicate more detailed explanations.

**Potential Impact:**

* Agents with longer messages may require summarization or training to be more concise.
* Consistent short messages might be beneficial for clarity, but could also mean less depth.



**4. Combined Word Cloud**

**Graph Description:**  
The word cloud highlights the most frequently used words across all processed messages, with larger text size indicating higher frequency.

**Insights:**

* **Key Topics:** The biggest words reveal the main conversation themes (e.g., “football,” “article,” “game,” “know,” etc.).
* **Repetitive Language:** Words that appear too frequently (e.g., filler words) might be targeted for more varied vocabulary or synonyms.
* **Context Clues:** If domain‐specific terms are large, it confirms that the conversation focuses on those topics (e.g., sports, politics, technology).

**Potential Impact:**

* Recognizing dominant keywords can guide topic modeling or content strategy.
* If certain words are unexpectedly large, it might suggest a repeated question or confusion area that needs addressing.



**5. Correlation Matrix**

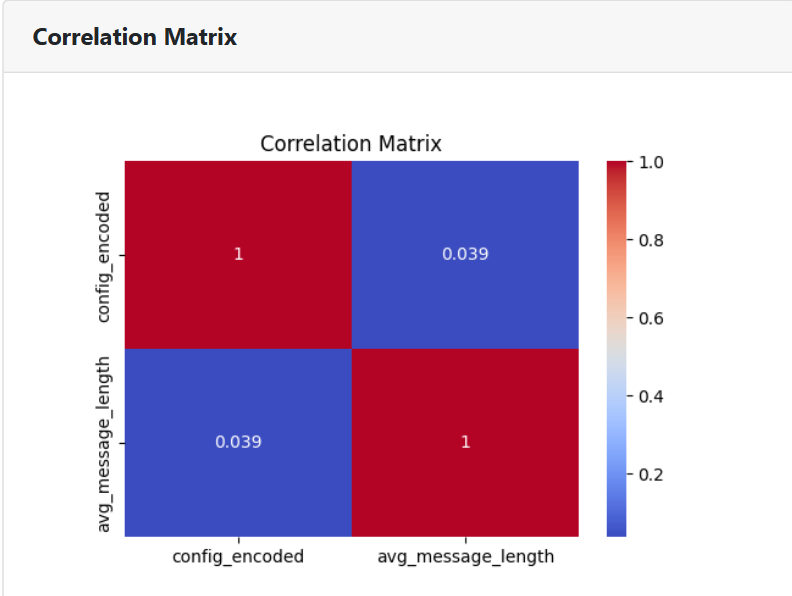
**Graph Description:**  
This heatmap displays the correlation coefficients among numeric features (for instance, config\_encoded and avg\_message\_length).

**Insights:**

* **Feature Relationships:** A strong positive correlation (closer to 1) between avg\_message\_length and config\_encoded might indicate that certain configurations (A, B, C, etc.) are associated with longer or shorter messages.
* **Weak or Negative Correlations:** Values near 0 or negative could show that two variables do not move together, or they move inversely.
* **Data Exploration:** If additional numeric features exist (e.g., conversation rating, turn rating), you can see if they correlate strongly with message lengths or config types.

**Potential Impact:**

* High correlation between features can guide further analysis (e.g., whether to combine or remove redundant features).
* Low correlation suggests independence, indicating that feature might provide unique insights.



**Overall Takeaways**

1. **Agent Dynamics:** The message count and sentiment distribution graphs reveal who talks more and how each agent feels or expresses themselves.
2. **Conversation Style:** The message length box plots suggest how detailed or concise each agent is.
3. **Topic Focus:** The word cloud highlights dominant keywords, indicating core discussion topics.
4. **Feature Interactions:** The correlation matrix shows relationships among numeric variables, helping refine modelling or deeper analysis.