OpenAI Dataset: Worker-Worker Dynamics and Diversity Metrics

Student Name: Hemish Savaliya
Enrollment ID: **202001206**
B.Tech. Project (BTP) Report
BTP Mode: ***On Campus***
Dhirubhai Ambani Institute of ICT (DA-IICT)
Gandhinagar, India
202001206[at] daiict.ac.in

Mentor's Name: Prof. Surish Dasgupta
Dhirubhai Ambani Institute of ICT (DA-IICT)
Near Indroda Circle Gandhinagar 382007, India
Sourish_dasgupta[at] daiict.ac.in

Abstract- **This project focuses on analyzing worker behavior through diversity metrics, utilizing the OpenAI dataset. A robust pipeline is developed to process worker and news ID interactions, compute embedding and category diversity, and cluster workers based on their behavioral patterns. The methodology combines key metrics, including Embedding Diversity and Category Diversity, into a unified Total Diversity score using a weighted formula, ensuring a balanced assessment of content and topic variation.**

The project highlights worker-level aggregation of diversity scores to identify behavioral trends and uses clustering techniques such as K-Means and DBSCAN to group workers with similar diversity patterns. Visualizations, including scatter plots and clustering maps, are employed to interpret and validate the results effectively.

This analysis provides insights into user interaction diversity and its implications for personalization and content recommendation. The scalable and modular pipeline serves as a foundation for future studies and applications in behavioral analysis and diversity-aware systems.

I. PROBLEM STATEMENT

Understanding worker behavior and diversity in content consumption is crucial for designing effective natural language processing systems and personalized content recommendations. This project focuses on analyzing diversity in worker interactions with news articles using the OpenAI dataset.

Key areas of focus include:

Computing diversity metrics such as embedding diversity and category diversity to measure content variation.

Combining metrics into a unified Total Diversity score for comprehensive analysis.

Aggregating diversity metrics at the worker level to understand individual behavior and provide a complete view of worker interaction diversity.

Clustering workers based on diversity scores to identify distinct behavioral patterns.

Visualizing diversity trends and patterns for actionable insights.

Developing a scalable framework for diversity computation and clustering.

This project provides a systematic approach to understanding worker diversity and interaction patterns, enabling advanced analysis of user behavior in the context of natural language processing and content recommendation systems.

II. INTRODUCTION

The OpenAI_final.ipynb notebook is designed to analyze worker behavior by computing diversity metrics and clustering users based on their interactions with news articles. The primary focus of this project is to evaluate the diversity of content consumption and topic exploration through the use of embedding-based and category-based diversity measures.

The pipeline processes worker and news ID data, consolidates it into actionable insights, and applies clustering techniques to group workers with similar behavioral patterns. The key metrics, Embedding Diversity and Category Diversity, are combined into a unified Total Diversity score to provide a comprehensive view of worker interactions. The diversity metrics are visualized and interpreted using tools such as scatter plots and clustering visualizations, which enable better understanding and validation of the results.

By aggregating diversity at the worker level and analyzing patterns across different user groups, this project provides valuable insights into user behavior. The scalable pipeline lays a foundation for extending the analysis to new datasets and tasks, facilitating further research in user behavior modeling and diversity-aware systems.

III. ABOUT THE DATASET

1. **Context :**
   The combined_openai dataset forms the backbone of this project, designed to analyze worker interactions with news articles and compute diversity metrics. The dataset represents a collection of worker actions, including interactions such as reading, summarizing, and categorizing news articles. By providing a structured view of user behavior, the dataset enables the computation of embedding diversity and category diversity, laying the foundation for clustering and behavioral analysis.
   The context of this dataset aligns with the broader goal of understanding user interaction patterns and diversity in the consumption of content. It plays a crucial role in the development of diversity-aware models and personalized recommendation systems.

2. **Content :**
   The dataset includes the following key components:

   - **Worker Information:** Unique identifiers for workers to track their actions and compute diversity metrics at the worker level.
   - **News ID:** Unique identifiers for news articles interacted with by workers, enabling the grouping of interactions and embedding computations.
   - **Article Embeddings:** Vectorized representations of news articles that allow the calculation of cosine similarity to measure embedding diversity.
   - **Category Embeddings:** Vectorized representations of article categories for calculating category diversity.
   - **Summaries:**User-generated summaries of news articles, reflecting worker engagement with the content.
   - **Confidence Scores:** Ratings provided by workers, indicating their confidence in the summaries they generated or their actions.
   - **Timestamps**: Temporal information about interactions, aiding in sequential analysis of actions.
   - Each row in the dataset captures a single worker-article interaction, enabling granular analysis of user behavior.

IV. TOOLS AND TECHNOLOGIES

I. **Libraries and Frameworks :**

- **Pandas:** Used for efficient data manipulation and analysis, including grouping, aggregation, and computation of diversity metrics.
- **NumPy:** Enabled numerical computations and matrix operations, particularly for working with embeddings.
- **scikit-learn:** Utilized for clustering algorithms (e.g., K-Means, DBSCAN) and data normalization.
- **Matplotlib:** Employed for generating scatter plots and visualizing clustering results to interpret diversity patterns.
- **TQDM:** Provided progress bars to track the status of computations, especially during diversity metric calculations.
- **cosine_similarity (from sklearn):** Used to compute similarity between embeddings, a key step in diversity computation.
- **Ast:** Facilitated parsing of string representations of embeddings into usable numerical arrays.

II. **Pretrained Models:**

- **BERT (Bidirectional Encoder Representations from Transformers):** Used for generating embeddings for news articles, summaries, and categories.Played a crucial role in capturing semantic relationships in textual data, enabling accurate computation of diversity metrics.

III. **Diversity Metrics:**

- **Embedding Diversity:** Calculated using cosine similarity between article embeddings to measure content variation.
- **Category Diversity:** Measured differences in topic exploration by analyzing category embeddings.
- **Weighted Total Diversity:** Combined embedding and category diversity metrics with custom-defined weights.

V. DIVERSITY FORMULA

➢ Formula Overview

To compute the **Total Diversity** for each worker, the following formula is used:

$$D_{total} = \alpha \cdot D_{embedding(norm)} + \beta \cdot D_{Category(norm)}$$

$$\text{Total Diversity} = \alpha \cdot \left(1 - \frac{\|v1\| \cdot \|v2\|}{v1 \cdot v2}\right) + \beta \cdot \left(1 - \frac{\|c1\| \cdot \|c2\|}{c1 \cdot c2}\right)$$

Where :

- $v1$ and $\vec{v2}$ are the embeddings of two consecutive articles.
- $\|v1^{\rightarrow}\|$ and $\|v2^{\rightarrow}\|$ are the magnitudes (norms) of these embeddings.
- $c1^{\rightarrow}$ and $c2^{\rightarrow}$ are the category embeddings (topics or genres) of the two consecutive articles.
- $\|c1^{\rightarrow}\|$ and $\|c2^{\rightarrow}\|$ are the magnitudes (norms) of these category embeddings.
- α and β are the weights assigned to embedding and category diversity, respectively.

➢ Component Defination :

1. **Embedding Diversity:**

   $D_{embedding}$ **= 1 - Cosine Similarity(**E_i,E_j**)**

   Where :

   - E_i,E_J: Embeddings of consecutive articles consumed by a worker.

2. **Category Diversity:**

   $D_{Category}$ **= 1 - Cosine Similarity(**C_i,C_j**)**

   Where :

   - C_i,C_J: Category embeddings of consecutive articles consumed by a worker.

3. **Normalization :**

   - Embedding diversity normalized:

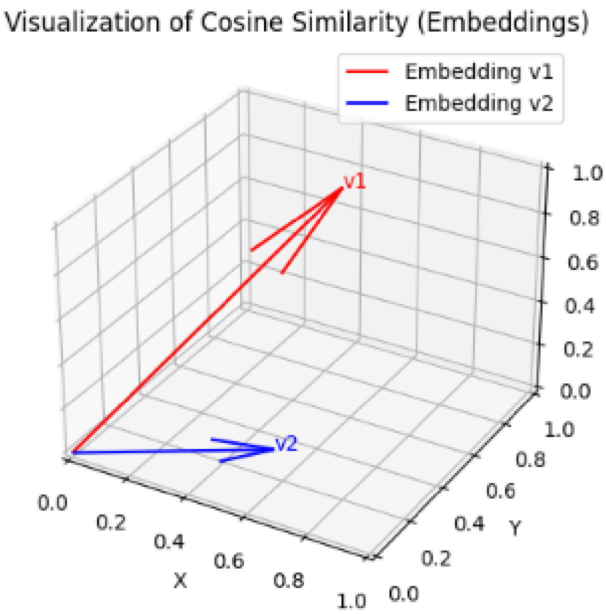
   $$D_{Embedding} = \frac{D_{Embedding}}{max(D_{Embedding})}$$

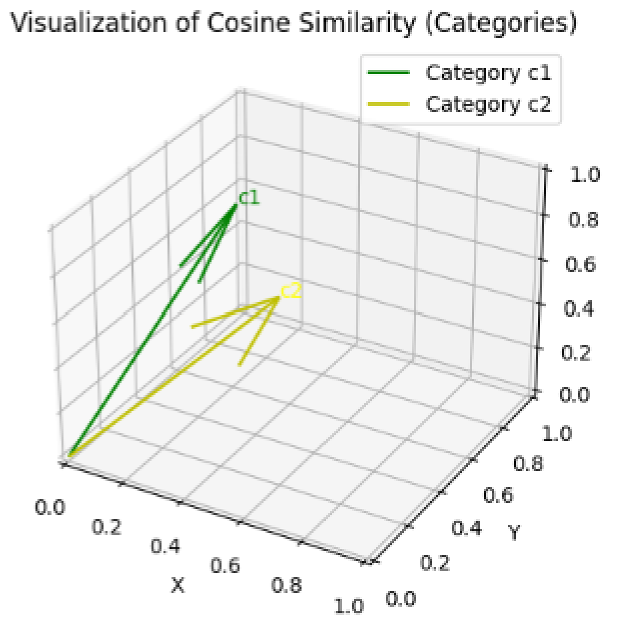
   - Category diversity normalized:

   $$D_{Category} = \frac{D_{Category}}{max(D_{Category})}$$

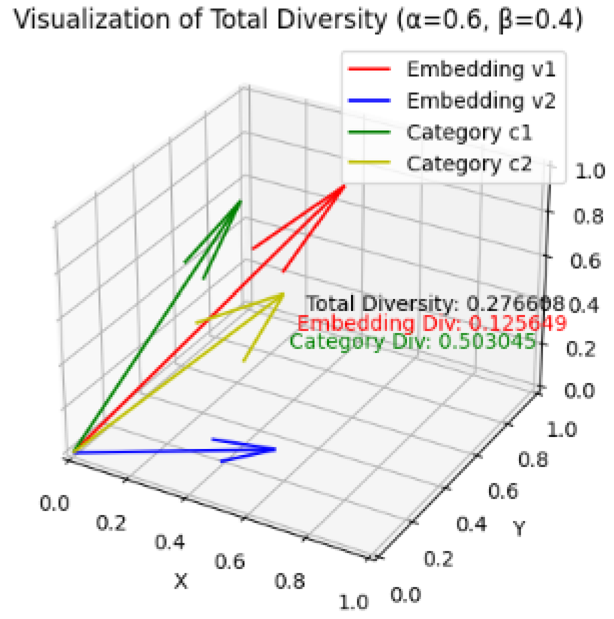
Purpose of the Formula

This formula combines **Embedding Diversity** and **Category Diversity** to provide a balanced view of user behavior, quantifying both content variation and topic exploration. The weights (α *and* β) enable flexibility, allowing prioritization based on specific application needs.





```
Diversity Calculation Outputs:
Alpha (α): 0.6, Beta (β): 0.4
Embedding Diversity (Normalized): 0.125649
Category Diversity (Normalized): 0.503045
Total Diversity: 0.276608
```



VI. IMPLEMENTATION DETAILS

Step by Step Explanation:

1. **Libraries and Setup**: Essential libraries like pandas, numpy, matplotlib, and tqdm are imported for efficient data manipulation, numerical computations, visualization, and progress tracking. These set up the foundation for all subsequent operations.
2. **Dataset Loading**: The project loads a structured dataset containing worker actions, news IDs, embeddings, and other metadata into pandas for efficient analysis.
3. **Worker and NewsID Grouping**: Data is grouped by worker and newsID to consolidate rows, simplifying data for further processing like embedding diversity computation.
4. **Worker-Level Aggregation**: Consolidates all data at the worker level, enabling the computation of diversity metrics and behavioral pattern analysis.
5. **Embedding Parsing**: Converts string representations of embeddings into numerical arrays for mathematical operations like similarity calculations.
6. **Pairwise Data Preparation**: Constructs consecutive newsID pairs for each worker to enable pairwise comparison for diversity calculations.
7. **Embedding Diversity**: Computes diversity as 1 minus the cosine similarity of consecutive article embeddings, capturing content variation.
8. **Category Diversity**: Similar to embedding diversity, it quantifies topic differences using category embeddings.
9. **Statistical Summaries**: Computes and analyzes minimum, maximum, and average diversity scores to detect trends or anomalies.
10. **Weighted Total Diversity**: Combines embedding and category diversity using predefined weights to provide a balanced diversity score.
11. **Worker-Worker Diversity**: Aggregates total diversity scores for each worker to understand overall behavioral diversity.
12. **Diversity Visualization**: Uses scatter plots and other visualizations to analyze and validate diversity patterns for specific workers.

13. **Worker Clustering**: Applies clustering algorithms (K-Means and DBSCAN) to group workers based on diversity patterns.
14. **Clustering Combination**: Integrates clustering results to provide robust and interpretable worker group assignments.

This modular pipeline integrates diversity metrics and clustering, enabling comprehensive behavioral analysis and actionable insights into user preferences.

VII. INSIGHTS

The analysis conducted in OpenAI_final.ipynb provides significant insights into worker behavior and diversity patterns, derived from the visualization of worker-level diversity using barplots and the accompanying table. These insights are critical for understanding individual preferences and informing strategies for personalized recommendations and optimized content delivery.

Barplot and Table Visualization:

1. **Worker-Level Diversity Representation**:
   - A barplot is used to represent the overall diversity scores for individual workers. Each bar corresponds to a worker, with the height reflecting their diversity score.
   - Additionally, a table provides a detailed breakdown of each worker's diversity metrics, including semantic diversity, categorical diversity, and overall diversity.
   - These combined visualizations provide a clear, unskewed representation of how workers vary in their content consumption preferences.
2. **Insights from Barplot and Table**:
   - Workers with **high diversity scores** engage with a wide range of content categories and topics, suggesting curiosity or broad interests.
   - Workers with **low diversity scores** demonstrate limited variation in content consumption, indicating a preference for specific topics or categories.
   - Workers with **moderate diversity scores** show balanced engagement, switching between exploratory and focused behavior.
   - The table further highlights how semantic and categorical diversity individually contribute to the final diversity score for each worker.
3. **Behavioral Patterns**:
   - The barplot and table together enable easy identification of behavioral patterns across workers, such as consistent niche behavior, broad exploratory tendencies, or a mix of both.
   - Unlike clustering-based approaches, these direct visualizations avoid skewed representation and provide an intuitive way to compare individual workers.
4. **Key Understandings:**

   Personalization: Diversity scores highlight individual preferences, allowing content recommendations to be fine-tuned for each worker.

   Exploration: High-diversity workers benefit from exposure to new and diverse topics, while low-diversity workers thrive with more focused and tailored content suggestions.

By replacing clustering with barplot visualizations and supplementing them with detailed tables, this analysis offers a more transparent and actionable understanding of worker behavior and diversity patterns.

VIII. CONCLUSIONS

This analysis provides a comprehensive and transparent approach to understanding diversity in worker behavior. By focusing on worker-level diversity metrics and using barplots for direct visualization, the project eliminates skewed representations and offers actionable insights. The integration of semantic and categorical diversity reveals how individual preferences vary across workers, enabling a clear understanding of exploratory and focused behaviors.

Key findings emphasize the value of tailoring content strategies to individual diversity patterns. High-diversity workers thrive with new and diverse content, sustaining their curiosity and engagement, while low-diversity workers benefit from precise, niche-focused recommendations that align with their preferences. The use of detailed tables complements visual insights, providing an in-depth breakdown of how diversity metrics contribute to overall behavior.

This refined methodology sets the foundation for future analyses of user preferences, offering a scalable framework that can adapt to diverse datasets and contexts, enhancing personalization and user engagement.