Smart Lead Gen: A Generative AI for Hyper-Personalized Professional Outreach

- Milestone 2: Dataset Preparation.
- Team: Group 7
- Course: DS and AI Lab
- Submission date: October 3rd, 2025

Datasets Used

1. (MBTI) Myers-Briggs Personality Type Dataset
2. Pandora - Big 5 Dataset
3. Enron Email Dataset

(MBTI) Myers-Briggs Personality Type Dataset
Link

This dataset offers a valuable resource for the development and fine-tuning of personality classification models. It comprises over 8,600 distinct samples, each containing two primary components: a labeled four-letter MBTI personality type and a corresponding corpus of the user's 50 most recent posts from the PersonalityCafe forum. This structure provides a direct mapping between an individual's written communication style and their classified personality profile, making it an ideal feature set for training a model to predict personality based on text analysis.

Pandora - Big 5 Dataset
Link

The Pandora dataset is a comprehensive collection of Reddit comments ideal for fine-tuning personality classification models. It contains data from over 10,000 users, with 1,600 of those users labeled with the well-established Big Five personality model. What makes this dataset particularly powerful is that it not only includes the raw text of user comments but also demographic information such as age, gender, and location. For the purposes of our project, the Big Five labels within this dataset will be reparametrized to correspond with the MBTI personality types. This strategic conversion will allow us to leverage PANDORA's extensive text and demographic data for the direct training and refinement of our MBTI-specific classification model, helping to create a more nuanced and accurate predictive tool.

Enron Email Dataset
Link

The sales message generation model will be fine-tuned using the Enron email dataset, a collection of approximately 500,000 emails exchanged within the Enron corporation. This dataset is an invaluable asset as it comprises real-world business communications, primarily from senior management. The inherently professional tone and structure of these emails provide a rich foundation for training our model. By learning from this corpus, the model will develop a nuanced understanding of corporate communication styles, appropriate language, and persuasive techniques used in a professional context. This will enable it to generate highly relevant and effective sales messages that resonate with a business audience.

Dataset Preprocessing

Email Message Generation

1. Keyword Filtering: Started by filtering the original, massive email dataset (containing over 500,000 entries) to create a smaller, more relevant subset. This was done by searching the subject and body of each email for a list of specific outreach and business-related keywords like "meeting," "proposal," and "opportunity."

2. Robust Text Cleaning: Developed a multi-stage cleaning function that was applied to the body of each filtered email to:

- Removed large, noisy blocks of text (like --- Forwarded by --- and ----- Original Message -----).
- Stripped out individual header lines (To:, From:, Subject:, etc.) that were mixed in with the body content.
- Isolated the clean, natural language of the core message, which was then saved into a new cleaned_body column.

3. Final CSV Creation: The final step was to prepare the output file. Selected the most important columns (subject, from, to, date, and our new cleaned_body) from the processed data and saved them into a final, ready-to-use CSV file named filtered_cleaned_enron.csv.

Personality Classification

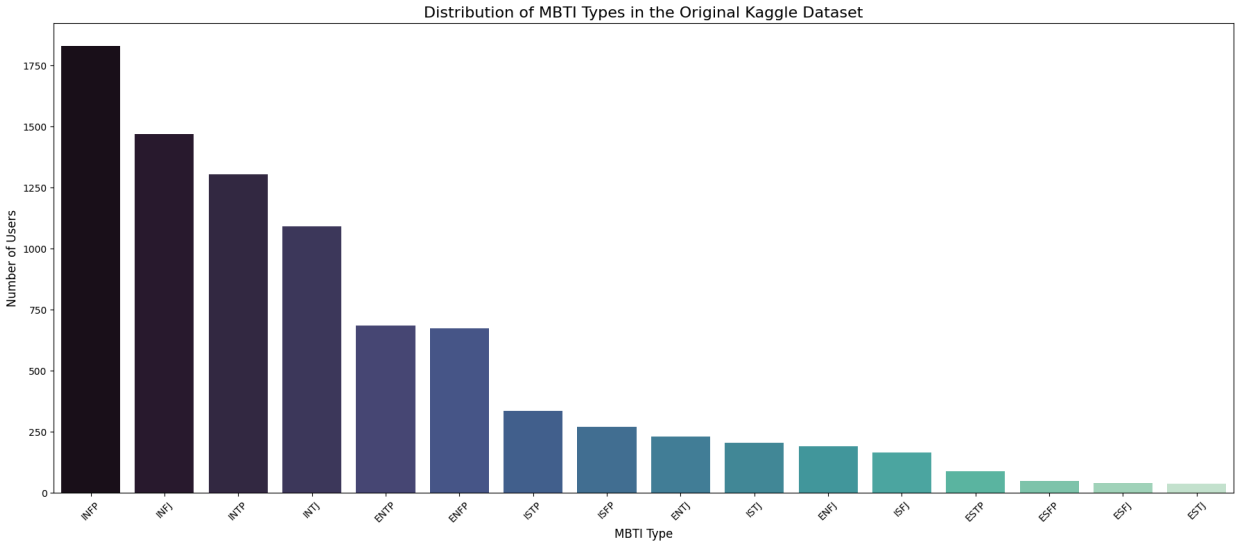
The distinct personality classification datasets, the MBTI personality type dataset and Pandora big 5 dataset have been concatenated into a unified corpus for analysis. The Pandora Big 5 classifications were processed and transformed into MBTI classifications based on each entry's MBTI dimensions: Introversion vs. Extroversion, Intuition vs. Sensing, Intuition vs. Sensing, Judging vs. Perceiving. The combined dataset was subjected to a multi-stage preprocessing pipeline:

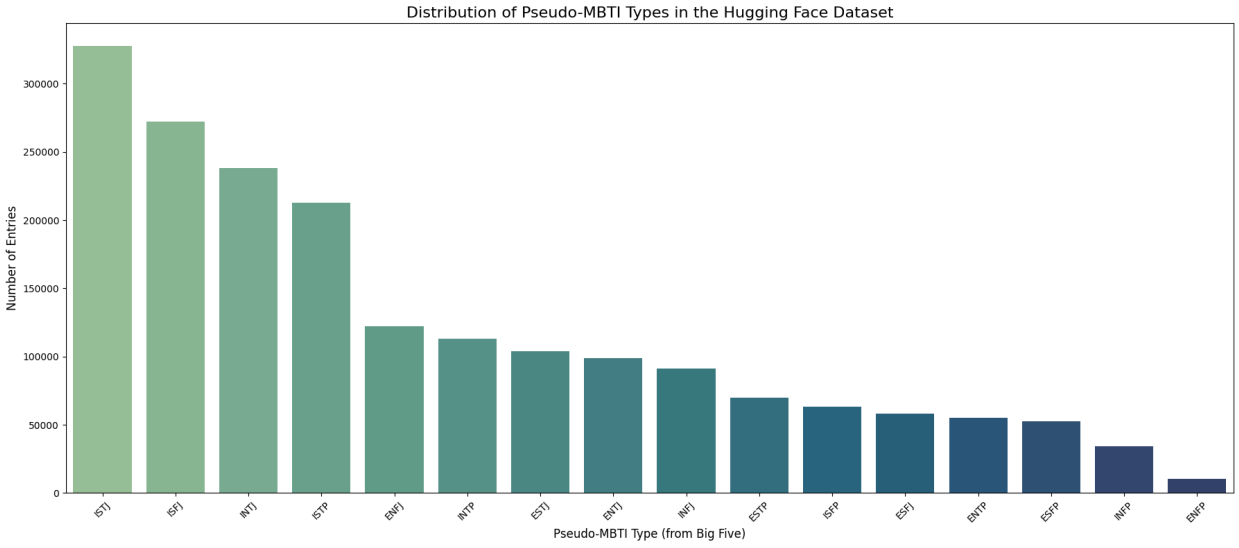
1. Data Restructuring: User posts, originally concatenated by '|||' delimiters, were parsed and expanded into individual entries. This transformed the data from a user-centric to a post-centric format.

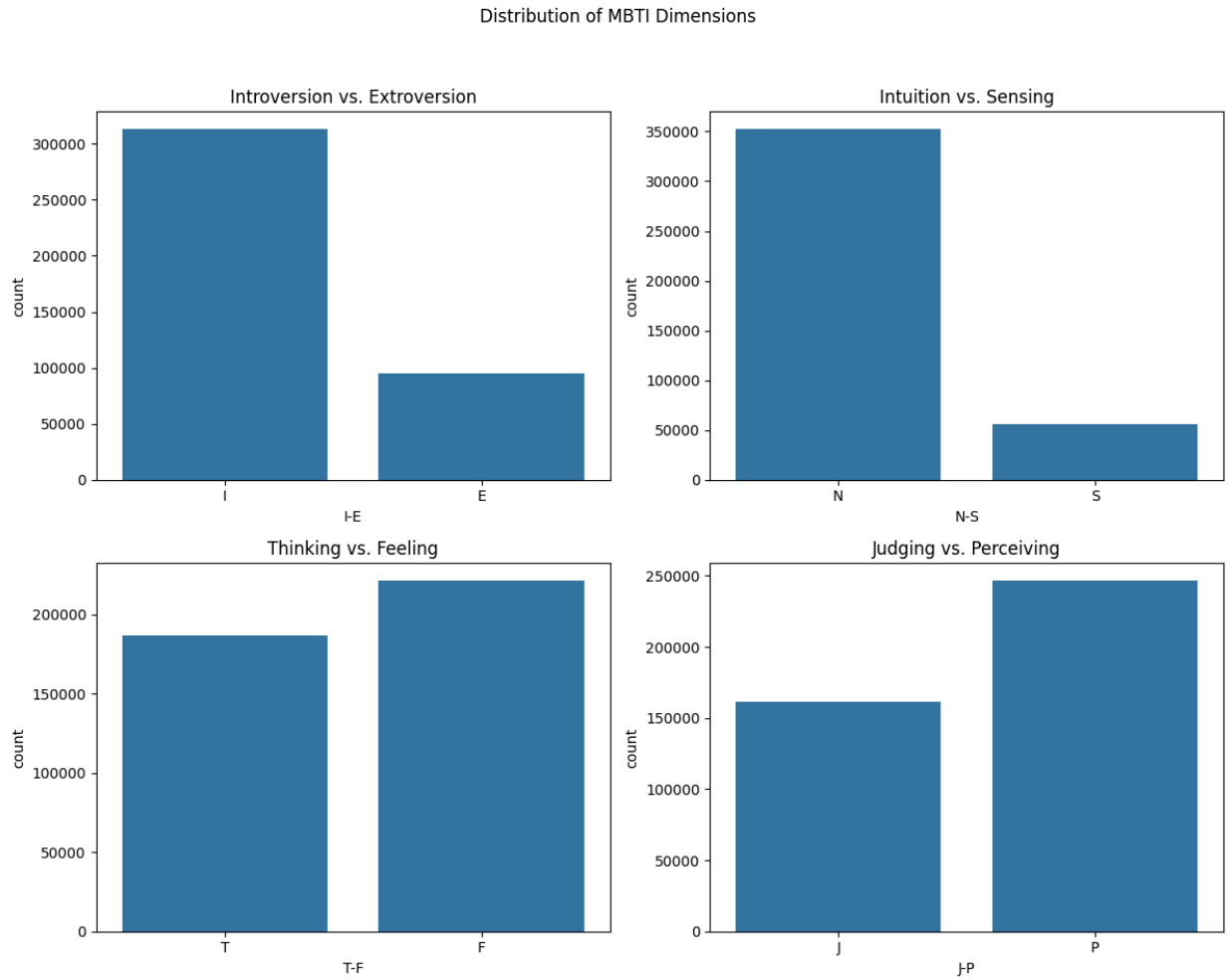
2. Feature Engineering: The 16-point 'type' label was deconstructed into four binary dimensional features: 'I-E', 'N-S', 'T-F', and 'J-P'.

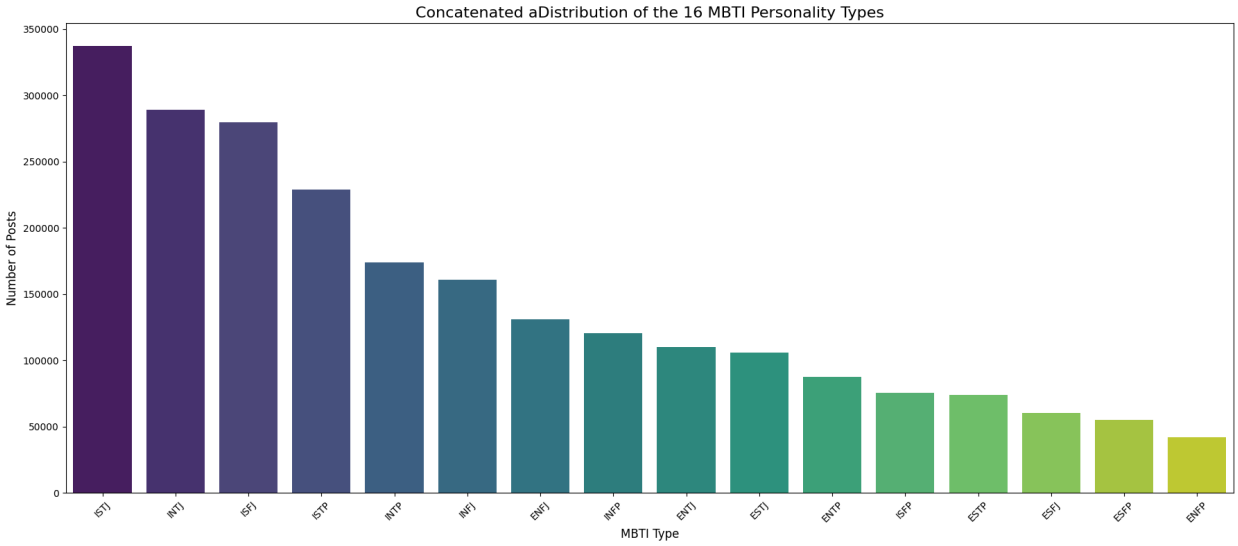
3. Text Normalization: A cleaning function was systematically applied to each text entry to remove URLs, punctuation, and other non-alphanumeric artifacts, ensuring data uniformity.

The final output is a clean, structured dataset ready for the feature extraction and modeling phase of the project. Listed below are a few EDA plots drawn to show the dataset distributions.






Distribution of MBTI Dimensions
Introversion vs. Extroversion
count
I
E
I-E
Intuition vs. Sensing
N
S
N-S
Thinking vs. Feeling
T
F
T-F
Judging vs. Perceiving
J
P
J-P


Concatenated aDistribution of the 16 MBTI Personality Types
Number of Posts
MBTI Type
ISTJ
INTJ
ISFJ
ISTP
INTP
INFJ
ENFJ
INFP
ENTJ
ESTJ
ENTP
ISFP
ESTP
ESFJ
ESFP
ENFP