**Automated Specific Response System for Email**

**MSDS692 – Data Science Practicum 1**

**Progress Report for Week 4**

**Project Details:**  
This week marked a crucial phase in my project, where I went deeper into my dataset through EDA and started laying the groundwork for my model with thorough feature engineering.

**Project Timeline:**

* Week 1: Initial project submission (DONE)
* Week 2: Data sourcing and project redirection (DONE)
* Week 3: Environment setup and data preprocessing (DONE)
* Week 4: Continued preprocessing and began initial model training. (DONE)

**Planned Work for the Week:**

My main objective was to treat my dataset through targeted preprocessing techniques and kickstart the model training process.

**Progress for the Week:**

* I break down and cleaned the 'Date' column in the dataset to remove time zone information, standardizing the format for easier analysis.
* Conducted an insightful EDA, identifying the most common email senders and analyzing the distribution of email body lengths, showing a majority of shorter emails.
* I extracted key terms from email subjects using TF-IDF vectorization, visually represented through bar charts and a word cloud, which shows the primary topics of discussion within the dataset.
* I did sentiment analysis of email bodies, categorizing them into positive, neutral, and negative sentiments. This analysis provided a foundational understanding of the emotional tone common in the communications.
* Begin the process of feature engineering, normalizing sentiment scores and preparing a feature set for model training.

**Roadblocks/Issues:**

* Faced challenges in correctly analyzing date formats and converting string data into a numerical format for model training, resolved through careful data preprocessing.
* The selection of features for effective model training triggered huge research and experimentation**.**

**Plan for Next Week:**

* Continue feature exploration and experimentation.
* Then focusing on developing logic for generating email responses based on model classifications.
* Begin testing various models with prepared feature set to identify the most effective approach for my classification needs.

**Resources for the Week:**

Used Pandas for data manipulation, Matplotlib and WordCloud for data visualization, Scikit-learn for feature engineering techniques, and NLTK for sentiment analysis.

**Project Timeline:**

Week 5: Develop response generation logic and test models.

Week 6: Model evaluation, maybe begin report drafting.

Week 7: Final model refinements, complete report draft.

Week 8: Finalize report and prepare presentation.