[Applied Analytics Project Week 3 Assignment]

Major: Applied Analytics

Name: Heng Bai, Yuang Guo

Review the Dataset Partition Strategy:

The dataset partition strategy used involves dividing the data into two key sets: a training set (twitter_training.csv) and a validation set (twitter_validation.csv). This approach is crucial because it allows for training models on a portion of the data while holding back the validation set to evaluate performance on unseen data. In sentiment analysis tasks like this one, where tweets are analyzed for sentiment, using a partition strategy helps ensure that the model can generalize beyond the training data. For example, if we have 10,000 tweets, 80% (8,000 tweets) might be used for training, while 20% (2,000 tweets) are held for validation. This ensures that the model isn't overfitted to the training set and performs well on unseen data, as evaluated on the validation set.

EDA Analysis and Insights:

Several exploratory data analysis (EDA) techniques were applied to the dataset to gain insights. One of the primary EDA steps involved generating a WordCloud to visualize the most frequently used terms in the dataset. For instance, common words like "good," "happy," and "great" might appear often in positive tweets, while words such as "bad," "sad," or "terrible" might dominate in negative tweets. Additionally, summary statistics showed the distribution of sentiment classes, and it was discovered that, for example, 60% of the tweets were labeled as "positive," 30% as "negative," and 10% as "neutral." This imbalance is important to address as it can lead the model to be biased towards the "positive" class. Furthermore, histograms and boxplots of tweet lengths revealed that most tweets have a length between 10-20 words, which suggests that the dataset has short and concise text, typical of Twitter.

Insights from EDA:

The EDA revealed several critical insights about the dataset. The class imbalance in sentiment distribution was one of the most important findings. For instance, if 60% of the tweets are positive and only 10% are neutral, this imbalance might cause the model to favor predicting the "positive" sentiment. Moreover, word frequency analysis showed that certain words appeared much more frequently in specific sentiment classes. Words like "love," "amazing," and "fun" were dominant in positive tweets, while negative tweets often included words such as "hate," "dislike," and "awful." This insight suggests that the vocabulary of the dataset is polarized based on sentiment, and feature selection should focus on distinguishing these frequent terms. Additionally, the data visualization showed that there are outliers in tweet length, with some tweets having unusually high word counts (e.g., more than 30 words), which might indicate special cases like promotional content or multiple hashtags that could affect sentiment prediction.

Data Problems and Preprocessing Recommendations:

The EDA uncovered several data quality issues that must be addressed through preprocessing. One of the most notable problems is the imbalance between sentiment classes. With a large percentage of tweets

falling into the "positive" category, the model could become biased and overpredict this class. To mitigate this, techniques like oversampling the minority classes (neutral and negative tweets) or using Synthetic Minority Over-sampling Technique (SMOTE) should be considered. Another issue is the presence of noise in the data, such as URLs, special characters, and hashtags, which can negatively impact text-based models. It is recommended to clean the text by removing these elements. Additionally, stop words (e.g., "the," "is," "in") should be removed as they do not contribute significantly to the sentiment analysis. Lastly, it might be beneficial to standardize the length of tweets by truncating or padding them so that all inputs to the model have a uniform length, ensuring that long tweets or very short tweets don't disproportionately influence the model.