**Data Collection and Preprocessing**

**Project Title :** Amazon Instruments Review

Summarizes the preprocessing steps applied to the Amazon Instrument Reviews dataset to ensure clean and consistent input for machine learning models.

**Preprocessing steps :**

| Section | Description |
| --- | --- |
| Data Overview | Loaded CSV dataset, checked missing values, and reviewed class distribution. |
| Handling Missing Values | Filled missing reviewText with empty strings and reviewerName with “ ". |
| Concatenation | Combined reviewText and summary into a single column reviews. |
| Labelling | Assigned labels: Positive (rating > 3), Negative (rating < 3), Neutral (rating = 3). |
| Text Cleaning | Lowercasing, punctuation removal, number removal, URL removal, newline removal. |
| Text Preprocessing | Tokenized reviews, removed stopwords, and applied WordNet Lemmatizer. |
| Polarity Scores | Calculated polarity of reviews using TextBOB for strength |
| Feature Engineering | Generated features: review length, word count, n-grams and dropped columns. |
| Encoding | Encoded target sentiment labels using LabelEncoder. |
| Vectorization | Applied TF-IDF with Unigram,Bigram, Trigram (max 5000 features). |
| Balancing | Used SMOTE to oversample minority classes. |

**Data Preprocessing Code Screenshots :**

| **Section** | **CODE** |
| --- | --- |
| Data Overview |  |
| Handling missing values |  |
| Concatenation |  |
| Labelling |  |
| Text Cleaning |  |
| Text Preprocessing |  |
| Polarity Scores |  |
| Feature Engineering |  |
| Encoding |  |
| Vectorization |  |
| Balancing |  |