Sentiment & Emotion Analysis – ML Module

Overview

This module focuses on **Machine Learning processing** for social media text data.
It performs **Sentiment Analysis** and **Emotion Detection** using **pretrained NLP models** in **Google Colab**.

The input data comes from the **Silver layer (Databricks)**, and the output is sent back to Databricks for **Gold-level analytics**

Objective

- Classify social media text into **sentiment** categories
- Detect **emotional tone** of each post
- Generate prediction confidence
- Produce ML-enriched data for analytics

Environment

- **Platform**: Google Colab
- **Language**: Python
- **ML Library**: Spark NLP
- **Compute**: Colab runtime (CPU)

Input Data

- Cleaned and prepared data from **Silver layer**
- Key columns:
  - `tweet_id`
  - `username`
  - `clean_text`
  - `created_at`
  - `hashtags`

Models Used

Sentiment Analysis

- **Model Type**: Pretrained Deep Learning Model
- **Output**:
    - Positive
    - Negative
    - Neutral

Emotion Detection

- **Model Type**: Pretrained NLP Emotion Model
- **Output Examples**:
    - Joy
    - Anger
    - Sadness
    - Fear
    - Surprise

ML Processing Steps

1. Load Silver data into Google Colab
2. Initialize Spark NLP pipeline
3. Apply sentiment prediction model
4. Apply emotion detection model
5. Extract predicted labels and confidence scores
6. Create final ML output dataset
7. Export results for Databricks Gold processing

Output Data

The ML process generates the following additional columns:

- `sentiment_label`
- `emotion_label`
- `confidence`

This enriched dataset is exported back to **Databricks** for Gold aggregation and dashboards.

Error Handling

- Try–except blocks for model execution
- Safe handling of null or empty text
- Controlled Spark session initialization

Limitations

- Uses pretrained models (no custom training)
- Batch processing only
- Limited compute resources (Colab free tier)

How to Run

1. Open the notebook in Google Colab
2. Upload or load Silver data
3. Install required NLP libraries
4. Run all cells in order
5. Export ML predictions

Role in Overall Architecture

- Acts as the **ML brain** of the pipeline
- Separates ML from data engineering
- Makes the system modular and scalable

Author

Yalini Sathiya.R Mathumitha.J
Databricks Certified Data Engineer Associate