Midterm Project Report: Reddit Sentiment & Emotion Analysis with AWS

🌍 Project Description:

I built a cloud-based sentiment and emotion analysis pipeline using Reddit posts as a data source. The goal is to extract user emotions from Reddit discussions, clean and transform the data, analyze sentiment and emotion with machine learning models, and visualize it via dashboards in Amazon QuickSight.

✈️ Tools & Services Used:

- **Reddit API (`praw`)** - to collect text data
- **Amazon S3** - storage for raw and processed data
- **AWS Glue Studio Notebook (PySpark)** - for cleaning and transforming text
- **SageMaker Notebook** - for applying sentiment and emotion models
- **Amazon QuickSight** - for dashboard visualization
- **HuggingFace Transformers** - local NLP models (fallback from Comprehend)

📊 Workflow Summary:

1. Data Collection

- Collected all posts from the `r/artificialinteligence` subreddit using Reddit API
- Number of rows: **495**
- Columns: title, body and createad_utc
- Stored raw posts in CSV format on S3

2. Data Cleaning with AWS Glue (PySpark)

- Combined `title` and `body` fields into one `full_text` field

- Removed URLs, special characters, and extra whitespace
- Tokenized the text (split into words)
- Removed stopwords using `StopWordsRemover`
- Dropped original columns, resulting in:
  - `full_text`: cleaned string
  - `filtered_words`: tokens with stopwords removed
- Saved cleaned data to (/cleaned_csv) but file parquet (because there are some problem with sagemaker csv)

3. Sentiment & Emotion Analysis (in SageMaker)

- Used **HuggingFace models**:

`distilbert-base-uncased-finetuned-sst-2-english` — a fine-tuned DistilBERT model for binary sentiment classification (POSITIVE / NEGATIVE)

`j-hartmann/emotion-english-distilroberta-base` — a DistilRoBERTa model fine-tuned for multi-class emotion classification (joy, fear, sadness, anger, etc.)

- Added columns:
  - `sentiment_label`, `sentiment_score`
  - `emotion_label`, `emotion_score`

4. Export and Upload

- Saved cleaned + analyzed data to CSV (`reddit_sentiments.csv`)
- Uploaded to `s3://aws-nlp-project/sentiment_output/`

5. Word Cloud Preparation

- Exploded `filtered_words` into individual rows
- Counted frequency and exported word-frequency CSV

- Created WordCloud locally using `wordcloud` and `matplotlib`
- Saved PNG to S3 and embedded it into QuickSight

6. Visualization with QuickSight

- Imported final CSV into QuickSight
- Created visuals:
    - Sentiment distribution (bar chart)
    - Emotion trends over time (line chart)
    - WordCloud (via S3 image)

🎯 Results & Insights

- Majority of posts are **negative** in sentiment (80%)
- Dominant emotions: `neutral`, `sadness`, and `surprise`
- WordCloud highlighted key discussion themes: "ai", "people", "model", "think", etc.
