PROGRESS REPORT WEEK 10

OVERVIEW

The main goal of these additions was to extract deeper emotional and affective insights from the lyrics using state-of-the-art Natural Language Processing (NLP) techniques.

To achieve this, we integrated the **CAMeL-Lab CAMeLBERT sentiment model**, a transformer-based model tailored for Arabic text, to accurately assess the sentiment polarity of each song. We also introduced a secondary layer of analysis to infer the **dominant emotion** in each song, offering a richer and more nuanced understanding of lyrical themes.

In addition to sentiment and emotion tagging, we implemented visualizations to uncover how these affective elements have evolved across different time periods. These plots not only add interpretability but also make it easier to communicate trends and patterns in the George Wassouf's musical journey.

SENTIMENT MODEL

1. We used the **CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment**

This BERT-based model understands Arabic morphology and syntax. It's designed to classify Arabic text into sentiment categories like Positive, Negative, and Neutral. It's effective because it is:

- Trained on both Modern Standard Arabic and dialects.
- Handles Arabic's rich structure better than generic multilingual models.

2. Sentiment and Emotion analysis

Sentiment Detection:
Using analyze_sentiment(), we parsed each song's preprocessed lyrics to extract:

- **Sentiment Label** (Positive, Negative, Neutral)
- **Sentiment Score** (quantitative measure of polarity)

Emotion Classification:
Building on sentiment, we then applied the infer_emotion() function. This uses both the sentiment and lexical content of lyrics to determine the **dominant emotion**— *joy*, *sadness*, *anger*, or *longing*.

Progress Visualization:
To keep users informed during large-scale processing, we enabled tqdm.pandas() for a clean progress bar while applying these functions row-wise to the dataset.

Output Saved:
The enriched dataset was saved as:

george_wassouf_lyrics_with_sentiment.csv

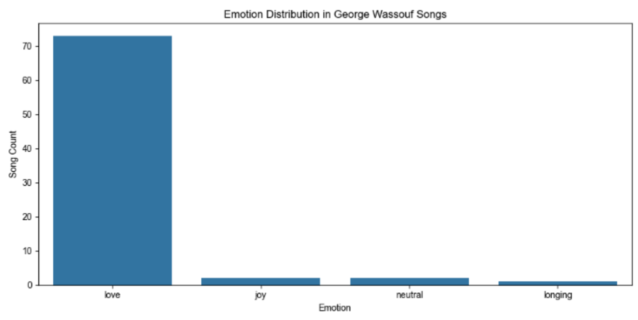
3. Visualisation

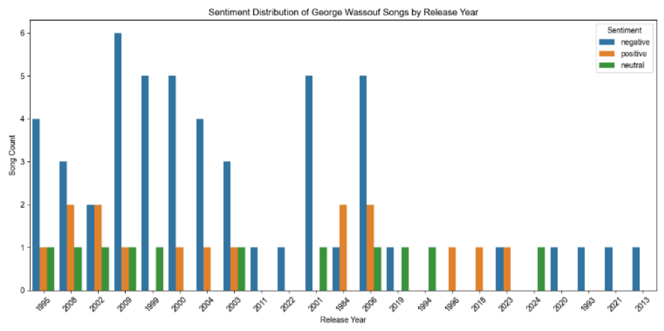
To make the new insights tangible, we visualized the results:

- **Sentiment Over Time**:
    - A seaborn count plot displays how sentiment varies by **release year**.
    - Great for spotting shifts in lyrical tone across decades.
- **Emotion Distribution**:
    - Another count plot showcases the overall distribution of **emotions** in the lyrics.
    - Gives a quick pulse check: Is Wassouf mostly melancholic? Joyful? Introspective?

These visuals are saved as:

- sentiment_by_year.png
- emotion_distribution.png





The emotion distribution plot showed us that we will need to work on emphasizing the other emotions more while training by increasing their weights and decrease love's probability

VECTOR EMBEDDINGS

4. Generating Vector Embeddings

We used the **AraBERTv2 model**, which is a general-purpose BERT model fine-tuned on Arabic. This model will be used to **generate vector embeddings** for entire songs, capturing the semantic meaning of lyrics beyond just sentiment.

5. Embedding Function

This function takes in a song's lyrics and returns a **768-dimensional embedding vector** using AraBERT. It:

- Tokenizes the lyrics.
- Passes them through the model.
- Extracts the hidden state (contextual embeddings).
- Averages across tokens (excluding special tokens) to get a single vector per song.

This creates a **numerical representation** of the song that can be used in clustering, similarity, and visualization.

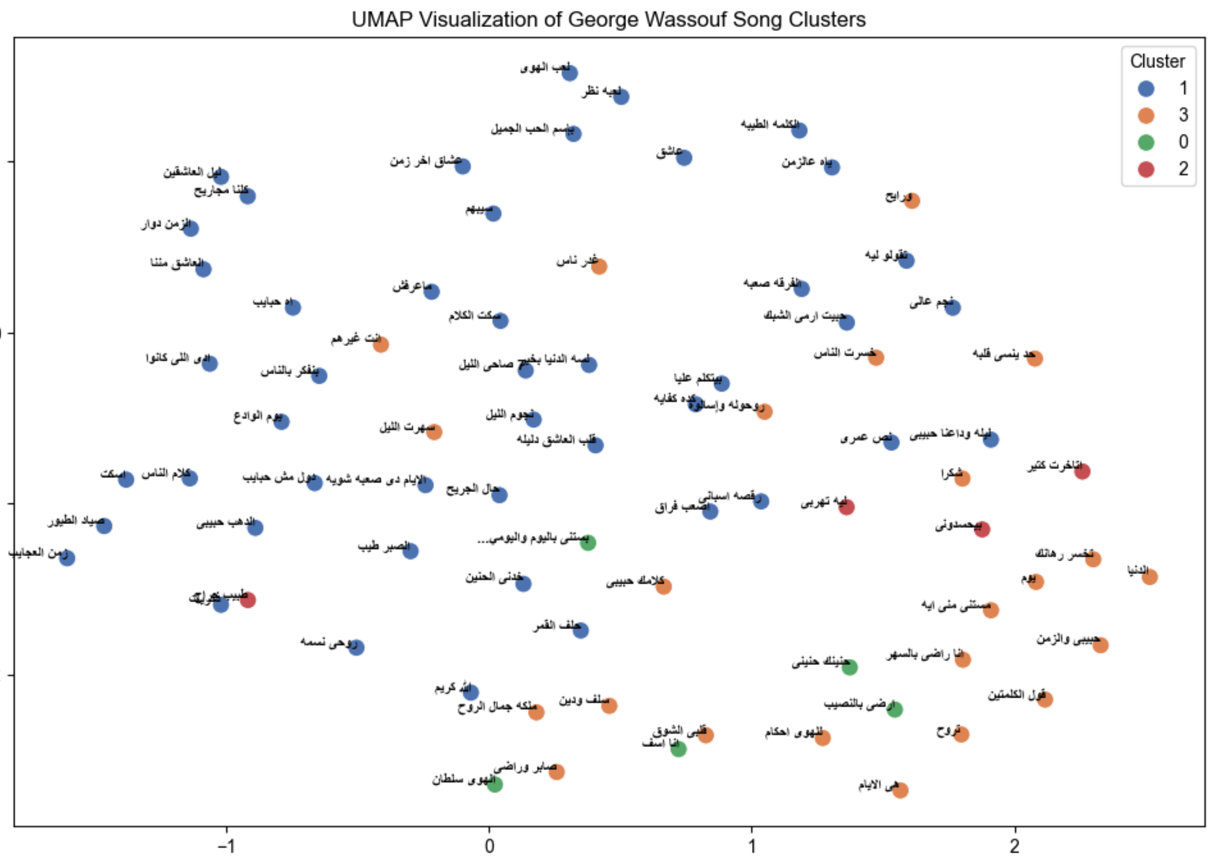
We applied the embedding function for all songs and saved the results in a new column

Output was saves a CSV file song_embeddings.csv

How it works:

1. Tokenization
   - The lyrics are broken down into subword tokens.
   - The tokenizer handles Arabic-specific morphology, such as prefixes (e.g., "ال") and suffixes.
   - It pads or truncates long songs to a maximum of 512 tokens, which is the BERT limit.
2. Passing Through Bert
   - The lyrics are passed through AraBERT.
   - Each token gets a **contextual embedding**,(it understands each word *in the context of the others*.)
3. Extracting the hidden states
   - This gives us a matrix of shape (1, sequence_length, 768).
   - Each token has a 768-dimensional vector.
4. Averaging token vectors
   - We remove special tokens like [CLS] and [SEP].
   - We average the vectors across all tokens to get **one vector per song**.
   - Why Average?
   - Averaging simplifies the data into one fixed-size vector.
   - It retains the overall theme and mood of the lyrics.

1. **Clustering**
   We group songs into **four clusters** based on their embeddings—intuitively grouping songs with similar lyrical themes (e.g., love, heartbreak, joy, longing). The clusters are saved as song_clusters.csv.
2. **Cosine Similarity**
   Calculates how similar each song is to others based on cosine similarity. For each song, the top 5 most similar songs are recorded and saved in song_similarities.csv. This can help identify thematic or stylistic resemblances across his discography.
3. **UMAP visualization**
   Uses **UMAP** (Uniform Manifold Approximation and Projection) to reduce the high-dimensional embeddings to 2D for visualization. Each song becomes a dot in 2D space, colored by cluster.
   Extra Touch: Arabic lyrics are reshaped and reordered using arabic_reshaper and python-bidi to display song titles correctly in the plot.


UMAP Visualization of George Wassouf Song Clusters
Cluster
🔵 1
🟠 3
🟢 0
🔴 2
لعب الهوى
لعبه نظر
بإسم الحب الجميل
الكلمه الطيبه
عشاق اخر زمن
عاشق
ياه عالزمن
ليل العاشقين
كلنا مجاريح
ورايح
سيبهم
الزمن دوار
غدر ناس
تقولو ليه
العاشق مننا
ماعرفش
الفرقه صعبه
اه حبايب
سكت الكلام
نجم عالى
حبيت ارمى الشبكه
انت غيرهم
خسرت الناس
حد ينسى قلبه
ادى اللى كانوا
بفكر بالناس
صاحى الليل
لمسه الدنيا بخير
بيتكلم عليا
كده كفايه
روحوله واسالوه
يوم الوادع
سهرت الليل
نجوم الليل
قلب العاشق دليله
نص عمرى
ليله وداعنا حبيبى
اتاخرت كتير
اسكت
كلام الناس
دول مش حبايب
الايام دى صعبه شويه
حال الجريح
شكرا
رقصه اسبانى
اصعب فراق
ليه تهربى
صياد الطيور
الدهب حبيبى
بيحسدونى
بستنى باليوم واليومي...
زمن العجايب
الصبر طيب
تخسر رهانك
الدنيا
خدنى الحنين
كلامك حبيبى
يوم
طبيب جراح
مستنى منى ايه
حلف القمر
روحى نسمه
حبيبى والزمن
حنينك حنينى
انا راضى بالسهر
الله كريم
سلف ودين
ملكه جمال الروح
ارضى بالنصيب
قول الكلمتين
قلبى الشوق
انا اسف
للهوى احكام
تروح
صابر وراضى
الهوى سلطان
هى الايام