

Week 11 Progress Report:

1. Emotion Analysis Using Translated Lyrics

We created a system to figure out the emotions in the songs, like joy, sadness, or anger, by translating the Arabic lyrics into English and analyzing them with a special tool.

- **Set Up the Tool:** We used a model called j-hartmann/emotion-english-distilroberta-base from the Hugging Face library. It can identify emotions such as joy, sadness, anger, fear, surprise, disgust, or neutral in English text.
- **Translation Step:** Since the lyrics are in Arabic, we used the GoogleTranslator from the deep-translator library to convert them to English so the emotion tool could understand them.
- **Emotion Detection:** We wrote a function called analyze_emotions_translated that:
 - Translates the Arabic lyrics to English and shortens them to 512 words to fit the model's limits.
 - Uses the emotion model to find the strongest emotion in the lyrics (e.g., sadness or joy).
 - Returns "neutral" if the lyrics are empty or if something goes wrong.
- **Visualization:** We made a bar chart showing how many songs each have emotion (e.g., how many are sad or joyful). This helped us see the emotional mix in George Wassouf's music.

This new method gave us a deeper look at the feelings in the songs, beyond just positive or negative.

2. Finding Important Words with TF-IDF

We used a technique called TF-IDF (Term Frequency-Inverse Document Frequency) to find the most important words in the lyrics and see what themes stand out. We applied it in two ways:

- **TF-IDF on All Lyrics:**
 - We used the TfidfVectorizer tool from scikit-learn to turn the cleaned-up lyrics (from the Tokenized & Preprocessed Lyrics column) into a set of numbers that show which words matter most.

- We found the top 25 words by adding up their TF-IDF scores across all songs. These words are the ones that are frequent in some songs but rare overall, making them special.
- To show Arabic words correctly in our charts, we used `arabic_resaper` and `bidl.algorithm` to fix their appearance.
- We created a bar chart of the top 25 words, which showed key themes in the lyrics, like love or longing.

- **TF-IDF by Composer:**

- We grouped all the lyrics by each composer and combined them into one big text per composer.
- We ran TF-IDF on these texts to find the top 10 words for each composer, revealing words that are unique to their songs.
- This helped us see if different composers have distinct styles or themes in their lyrics.

These TF-IDF results are a starting point for understanding the main ideas in the songs and how composers differ.

3. Improving Sentiment Analysis

We worked on two ways to measure if the songs feel positive, negative, or neutral, giving us two perspectives on the lyrics' mood:

- **Simple Word-Based Sentiment:**

- We made a function called `sentiment_score` that counts positive words (like "حب" for love or "فرح" for joy) and negative words (like "حزن" for sadness or "ألم" for pain) in the lyrics.
- The score is calculated by subtracting the number of negative words from positive ones and dividing by the total words, so we get a value showing how positive or negative the song is.
- We applied this to the cleaned lyrics and created a chart:
 - A line showed the average sentiment for each year the songs were released.
 - Dots showed the sentiment of individual songs, with a line at zero to mark neutral.

- **Advanced BERT-Based Sentiment:**

- We used the CAMEL-Lab/bert-base-arabic-camelbert-da-sentiment model, which is trained to understand Arabic and classify text as positive, neutral, or negative.
- Our sentiment_score function turns the model's results into a score: positive adds to the score, negative subtracts, and neutral adds nothing, giving a value between -1 and 1.
- We made a similar chart with a line for average sentiment per year and dots for individual songs.

What's Next

1. Make TF-IDF Better:

- Try different TF-IDF settings, like looking at pairs of words or removing common Arabic words (stop words), to find more meaningful terms.
- Use tools like PCA or t-SNE to turn TF-IDF results into visuals that show how songs are similar or different.
- Group songs into clusters (e.g., using K-Means) based on TF-IDF to find songs with similar themes.

2. Refine Sentiment Analysis:

- Add more positive and negative words to the word-based method, including regional or poetic terms.
- Adjust the BERT model to better handle short or tricky lyrics, maybe by training it more on our data.
- Compare the two sentiment methods to see if they agree and which works best.

3. Generate Lyrics:

We plan to explore a pretrained language model, such as an Arabic-specific transformer, and fine-tune it on the George Wassouf lyrics dataset to better capture the stylistic and thematic nuances of his songs. This fine-tuned model will be used to generate new lyrics in the style of George Wassouf, allowing us to create original content that reflects his emotional and linguistic patterns.