Deep Learning Project

Lyrics Analysis on Amr Diab’s Songs

Nour Atef (8129) - Marly Magued (8052) - Arwa Moustafa (8278) - Esraa Saleh (8162)

---

1 Introduction

Fig. 1 *Amr Diab*

Music has long been a mirror of cultural expression, and lyrics in particular offer a unique window into the emotional and social landscapes of their time. In recent years, lyric analysis has emerged as a compelling interdisciplinary field, blending linguistics, cultural studies, and data science to uncover patterns in language, sentiment, and thematic content. This project builds on that growing interest by turning its focus to the iconic Egyptian artist Amr Diab, whose expansive discography offers a rich canvas for lyrical exploration.

Known affectionately as "El Hadaba," Amr Diab has redefined Arabic pop music through a career that spans more than forty years. Blending classical Arabic motifs with global pop, dance, and electronic influences, his work transcends borders and generations. His songs are not only celebrated for their catchy melodies but also for the emotional depth and narrative range found in his lyrics, whether they speak of love, longing, celebration, or introspection. With multiple international awards and a deeply loyal fanbase, his influence on the music industry is both enduring and profound.

In this study, we aim to explore Amr Diab’s lyrical world through the lens of computational analysis. By applying natural language processing (NLP), sentiment analysis, and clustering algorithms to a curated dataset of his lyrics, sourced from Genius.com, we seek to uncover hidden trends, recurring themes, and emotional arcs across different periods of his career. This approach allows us to go beyond surface-level appreciation and instead gain a more nuanced understanding of how his music evolved in tandem with cultural shifts and personal expression.

Ultimately, this project is not just a technical exercise, but an attempt to bridge data and art, using numbers and models to better understand the heart behind the music. It offers a small contribution to the growing dialogue between technology and the humanities, while celebrating the legacy of an artist who has left an indelible mark on the soul of Arabic music.

2 Data Collection

The dataset used in this case study was compiled from Genius.com, a popular website for song lyrics. We focused specifically on Amr Diab's discography, gathering a comprehensive list of songs alongside their corresponding lyrics. The data was then saved as a CSV file, as shown in *fig.2*, for ease of processing and analysis. Prior to using the dataset, a thorough inspection was carried out to ensure consistency and accuracy, particularly focusing on resolving any inconsistencies in the lyrics themselves.

The dataset includes several key attributes: the song's title, year of release, the composer's name, the lyricist's name, and the full lyrics of the song. These attributes form the foundation for our analysis, which aims to uncover trends and insights related to Diab's music over time.

	Year	Composer	Lyricist	Song	Lyrics
0	2023	محمد أحمد فؤاد	تامر حسين	بيوحشنا	...ملازn\ملازمنا، ملازمنا خياله وطيفه فين ما نروح
1	2023	أحمد إبراهيم	أيمن بهجت قمر	معرفش حد بالأسم ده	...أنا اللي تاه عقله ولقاهn\ما أعرفش حد بالإسم دا
2	2023	محمد يحيي	بهاء الدين محمد	ظبط مودها	... وأؤمرn\أطلب حتى عينيها تاخدهاn\لما تظبط مودها
3	2023	محمد يحيي	محمد القاياتي	سلامك وصلي	...وأتاريني واحشك زي ما إنت واحشني ياn\سلامك وصلي
4	2023	محمد يحيي	محمد البوغة	واخدين راحتهم	...n\واخذين راحتهم قاعدين في قلبي مربعين وبيعصروه
...	...	...	...	...	...
305	1983	هاني شنودة	هاني ذكي	الزمن	...اللي كn\الزمن بينسى دايماً، مع الزمن مفيش وعود
306	1983	هاني شنودة	عبد الرحيم منصور	نور يا ليل	...\يا اللي عشقناك وإحنا صغارn\نور يا ليل الأسرار
307	1983	عزمي الكيلاني	عصام عبدالله	وقت وعشناه	...n\وقت وعشناه إنتي وأنا، جرح حفرناه لبقية عمرنا
308	1983	ياسر عبد الحليم	عوض الرخاوي	أحلى دنيا	...\مالية شفايف كل الناسn\إمتى نشوف البسمة الحلوة
309	1983	هاني شنودة	هاني ذكي	أحضان الجبل	...في يوم والشمس طالعة بألوانها الرقيقة ولمستها ا

Fig. 2 *Collected Dataset of Amr Diab's Songs*

3 Data Preprocessing

Data preprocessing is crucial for ensuring that raw data is clean, consistent, and in a format suitable for analysis. It improves the accuracy of models by eliminating noise, handling missing values, and addressing outliers. Preprocessing also helps standardize and normalize data, which is vital for achieving consistent results across different datasets and models. By reducing errors, simplifying complexity, and enhancing data quality, preprocessing leads to more reliable and efficient analysis, ultimately boosting the performance of machine learning models and increasing the overall success of data-driven decision-making.

In this study, we utilized several specialized libraries to handle and preprocess Arabic text. The primary library used for text processing was Camel Tools, which is tailored specifically for Arabic natural language processing (NLP). Camel Tools offers linguistically informed preprocessing functions that address the unique characteristics of the Arabic language, making it a better fit than general-purpose NLP libraries. The preprocessing steps included tasks such as Unicode

normalization, diacritic removal, and tokenization, which are crucial for effectively analyzing Arabic text.

3.1 Reading the Dataset

The lyrics dataset was read into a Pandas DataFrame from the CSV file, making it easy to manipulate and preprocess. In addition to the song data, we also loaded a list of stopwords that are commonly found in Arabic text. These stopwords, which include words like “و” (and), “في” (in), and “على” (on), were removed from the lyrics to reduce noise and improve the performance of the analysis.

To ensure the dataset was clean and ready for analysis, we first inspected it for any missing values. We used the `dropna()` function to remove rows with missing data, ensuring that no null values were present in the final dataset. This is an important step in data cleaning, as missing data could lead to errors or inaccurate analysis.

3.2 Removing Stop Words

One of the first steps in text preprocessing is to remove stopwords, which are commonly occurring words that do not contribute significant meaning to the content. For example, in the sentence " أنا الذي أحبك" (I am the one who loves you), words like "أنا" (I) and "الذي" (who) are stopwords, while "أحبك" (love you) is the meaningful word.

We created the function `remove_stopwords()` to remove stopwords from each song's lyrics, allowing us to focus on the more meaningful terms in the text. This function was applied to each song's lyrics, significantly reducing the noise in the data and preparing the text for further analysis.

3.3 Cleaning Text

Next, we implemented the `clean_text()` function to perform additional preprocessing on the lyrics. The function includes several important steps:

1. Removal of English characters, numbers, punctuation, and extra spaces.
2. Unicode normalization to standardize the character encoding.
3. Orthographic normalization to standardize variations in Arabic letters (e.g., converting "أ" to "ا").
4. Removal of diacritics, which are marks that appear above or below letters and can vary based on dialects or stylistic choices.

After cleaning, the text was significantly more consistent, with extraneous elements like English characters and numbers removed, and Arabic text normalized for analysis.

3.4 Tokenization

Tokenization is the process of breaking down text into smaller units, such as words or phrases. This step is essential for transforming the raw text into a structured form that can be analyzed. Camel Tools' `simple_word_tokenize()` was used to tokenize the Arabic lyrics.

3.5 Text Processing

Arabic text processing poses unique challenges due to the language's rich morphology and complex script. Arabic words change depending on tense, gender, case, and the presence of diacritics. For instance, words like "علَم" (flag) and "عِلْم " (knowledge) are distinguished only by diacritics, and variations in the letter "Alef" (ا, أ, إ) can change the meaning of words. Text preprocessing helps address these challenges by normalizing these variations and reducing ambiguity.

We applied the function `find_non_normalized()` to identify non-normalized characters and diacritics in the lyrics to ensure that our text was consistently processed.

3.4 Categorizing Decades

The final step in our analysis involved categorizing the songs by decade. This categorization enables us to examine trends in Amr Diab's music over time, such as shifts in themes, sentiment, or musical style. By analyzing the lyrics across different periods of his career, we can gain a better understanding of how his artistic approach has evolved and how it reflects broader cultural and social changes. This analysis will provide a richer context for appreciating Diab's enduring influence on the Arabic music scene.

	Year	Composer	Lyricist	Song	Lyrics	Decade	Composer_first_name	Composer_last_name	Lyricist_first_name	Lyricist_last_name
0	2023	محمد أحمد فؤاد	تامر حسين	بيوحشنا	[ملازمنا, ملازمنا, خياله, وطيفه, فين, نروح, مل...	Early 2020s	محمد	أحمد فؤاد	تامر	حسين
1	2023	أحمد إبراهيم	أيمن بهجت قمر	معرفش حد بالأسم ده	[اعرفش, بالاسم, تاه, عقله, ولقاه, اعرفش, بالاس...	Early 2020s	أحمد	إبراهيم	أيمن	بهجت قمر
2	2023	محمد يحيي	بهاء الدين محمد	ظبط مودها	[تظبط, مودها, اطلب, عينيها, تاخدها, واوْمر, واج...	Early 2020s	محمد	يحيي	بهاء	الدين محمد
3	2023	محمد يحيي	محمد القاياتي	سلامك وصلي	[سلامك, وصلي, واتاريني, واحشك, واحشني, وهتفضل...	Early 2020s	محمد	يحيي	محمد	القاياتي
4	2023	محمد يحيي	محمد البوغة	واخدين راحتهم	[واخذين, راحتهم, قاعدين, قلبي, مربعين, وبيعصرو...	Early 2020s	محمد	يحيي	محمد	البوغة

Fig. 3 *Some Amr Diab's Songs After Processing*

4 Lyrics Analysis

With the foundational data preprocessed and structured, we now turn our attention to a deeper exploration of the lyrical content. This next phase involves analyzing the lyrics themselves to uncover the underlying themes and deeper narratives beyond surface-level listening. Unlike acoustic or musical analysis, lyrics analysis provides a direct window into the songwriter's intent and the story conveyed through language. In our study, we conduct this analysis to gain insights into the

content of Amr Diab's lyrics and enriching our understanding of his artistic evolution and the emotional resonance of his work.

4.1 High Level Statistical Aggregates of the Data

In order to gain a foundational understanding of the lyrical dataset, we began our analysis by computing several high-level statistical aggregates. These statistics offer an overview of the scope and diversity within Amr Diab's song collection. We found that the dataset comprises 309 unique songs, reflecting a substantial body of lyrical work. An examination of song credits revealed contributions from 77 distinct composers and 55 lyricists, suggesting a notable degree of collaboration and artistic variety over the years. To further contextualize the lyrical richness, we assessed the total number of words used across all songs after removing Arabic stopwords, arriving at 38,024 words in total. Moreover, we identified 5,872 unique words within the corpus, indicating a considerable lexical diversity that invites deeper exploration through subsequent thematic, sentimental, and emotional analysis. These aggregates provide essential insight into the scale and texture of Amr Diab's lyrical expression.

4.2 Analysis to Specific Words

To explore the emotional and thematic depth of Amr Diab's lyrics, this section focuses on a set of specific, high-impact words that frequently recur throughout his discography. By analyzing individual keywords, we aim to quantify the lyrical weight of central themes and understand how they contribute to the emotional fabric of his music. The first and most prominent of these is the word "حب" (love), which forms the cornerstone of many of his songs and serves as a gateway to exploring the broader patterns in his lyrical expression.

4.2.1 Keyword حب (Love)

This section plays a crucial role in highlighting the thematic core of Amr Diab's lyrical work by focusing on the keyword "حب" (love), a central and emotionally resonant concept in Arabic music. Given Amr Diab's reputation as a leading voice in romantic Arabic pop, analyzing how often and in what forms the word "حب" appears offers valuable insight into both the lyrical emphasis and emotional tone of his songs. This type of targeted keyword investigation helps quantify how deeply love permeates his discography, supporting broader claims about his musical identity and audience appeal.

The analysis began by filtering all lyrics that contain any word featuring the subword "حب" (love), including variations like "حبك" (your love), "أحب" (I love), or "محبة" (affection). Results revealed that such words appeared 2,209 times across all lyrics, which amounts to 5.81% of the total words, clearly showing high lexical emphasis. When considering only unique words, 129 different love-related words were found, representing 2.20% of all unique vocabulary used.

Furthermore, the presence of this theme spans broadly across the artist's body of work: out of 309 total songs, 244 songs (or 78.96%) include some form of the word "حب", affirming its importance at the song level, not just in word count, but in narrative focus. A list of these songs showcases the

extent of love’s presence across decades and musical styles in Amr Diab’s catalog. This analysis confirms that love is not merely a lyrical decoration, but a foundational theme deeply embedded in his artistic expression.

Finally, a frequency-based breakdown of love-related words per song was compiled to identify which songs emphasize the theme the most. A smaller, random sample of these songs was selected for closer inspection, allowing for a more qualitative look at how the concept of love is lyrically framed and repeated.

4.2.2 Comparing with Other Words Across Different Themes

Following the detailed analysis of the word "حب" (love)**,** we expanded the investigation to include other emotionally charged and thematically significant words in Amr Diab's lyrics. This comparison allows for a broader understanding of the lyrical focus and emotional range present in his discography. Using the same methodology, we calculated the frequency and song-level coverage of keywords like "قلب" (heart), "عين" (eye), "دموع" (tears), and "حياة" (life)**.** The results reinforced the dominant role of romantic themes: love remains the most prevalent word, appearing in nearly 80% of all songs and making up over 5% of the total vocabulary. "Heart" followed closely, also strongly present in more than two-thirds of his songs, further emphasizing emotional expression. In contrast, "eye" appeared in about a third of the songs, typically symbolizing longing or beauty. Interestingly, the word "tears" was rare, used sparingly to convey sadness, and the word "life" did not appear at all, suggesting that Amr Diab's lyrical universe is more focused on personal, intimate emotions than on abstract reflections about life. This thematic analysis provides strong evidence of Amr Diab's consistent commitment to romantic and emotional expression across his musical career.

4.3 Temporal Rate of Singing Performance

This subsection analyzes the temporal dynamics of Amr Diab’s musical output, focusing on how many songs he released per year and per decade. By plotting the number of songs produced annually between 1983 and 2023 (see *fig. 4.3*), we observe notable trends in his artistic activity over time. The 1990s and 2000s, especially, stand out for their steady production rates, reflecting a period of creative consistency and possibly peak popularity. Certain years, such as 2020, saw unusually high output, likely corresponding to the release of major hits. Interestingly, there was a dip in song releases around the early 2010s, which may be attributed to socio-political factors affecting Egypt during that time. Overall, this temporal analysis reveals both the endurance and fluctuations in Amr Diab’s career trajectory, offering a clearer picture of his long-term artistic evolution.

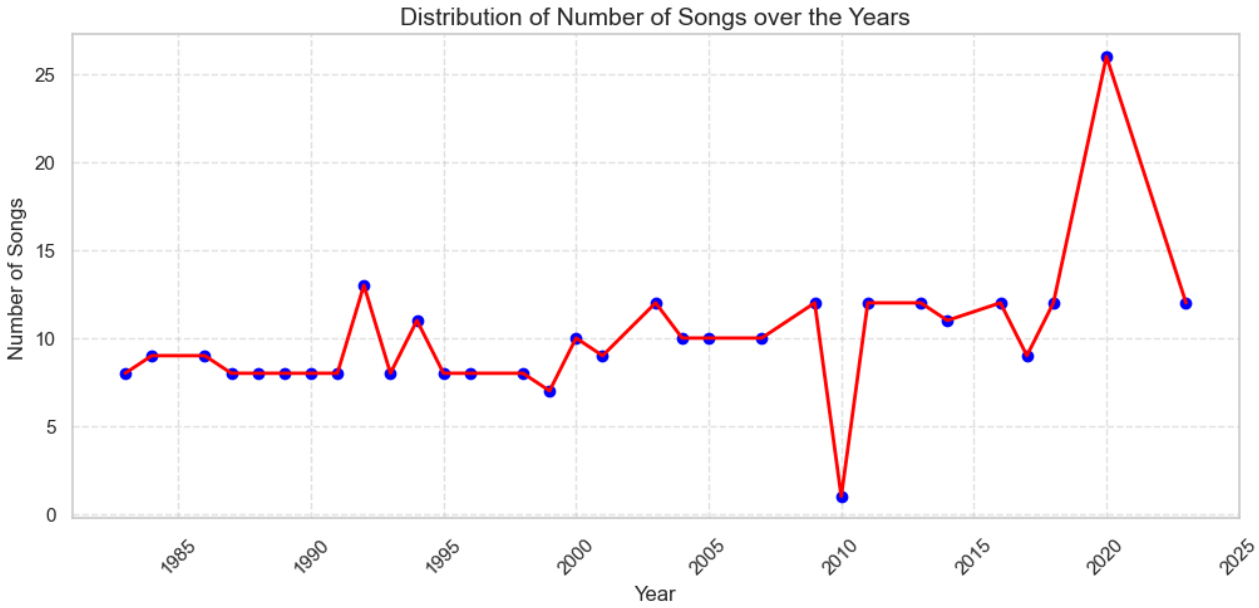

Fig. 4.3 *Line Graph of Number of Amr Diab's Songs over the Years*

4.4 Text Mining

Text mining offers a powerful lens for exploring lyrical content by uncovering hidden patterns and thematic structures within language. Using Natural Language Processing (NLP) techniques, we can systematically analyze the complexity of song lyrics beyond surface-level impressions. This includes examining how often certain words are used, how long they are, and how diverse and dense the vocabulary is. Such measures, collectively referred to as *lexical complexity*, provide insight into the lyrical style and evolution of an artist. In the case of Amr Diab, whose discography spans decades and genres, analyzing lexical complexity allows us to track how his songwriting has developed in relation to cultural shifts, technological advances, and audience preferences.

4.4.1 Word Frequency

Word frequency serves as one of the core indicators of lyrical complexity and plays a key role in shaping the memorability and emotional impact of a song. In this subsection, we explore how often words appear in Amr Diab's lyrics, not just unique words, but the total number of words used in each song. By identifying songs with the highest and lowest word counts, we gain insight into changes in musical composition over time. For example, we observe that many of Amr Diab's longest songs, lyrically, were released during the 1980s, possibly reflecting the influence of live performances and longer orchestral arrangements prevalent at the time. In contrast, more recent songs tend to be shorter, aligning with modern consumption habits shaped by fast-paced digital media and shorter attention spans. A comparative analysis of the 1980s and 2000s further highlights a notable shift in lyrical length, offering a quantitative view of how songwriting trends have evolved across decades.

Furthermore, by calculating the number of words in each song's lyrics and creating a histogram overlaid with a kernel density estimation (KDE) curve (see *fig 4.4.1.1*), we were able to visualize the

overall distribution. From the analysis, the average number of words per song is 123, with a median of 116 and an estimated mode of 107 words. These statistics reveal that while most songs are relatively short, the average song length is slightly longer due to a few outliers with more words. The right-skewed distribution indicates that many of Amr Diab's songs are concise, but longer tracks are still present, pulling the average upwards.

This right-skewed distribution suggests that most of the songs fall in the 100–150 word range, but longer songs (often exceeding 200 words) are pushing the mean higher.

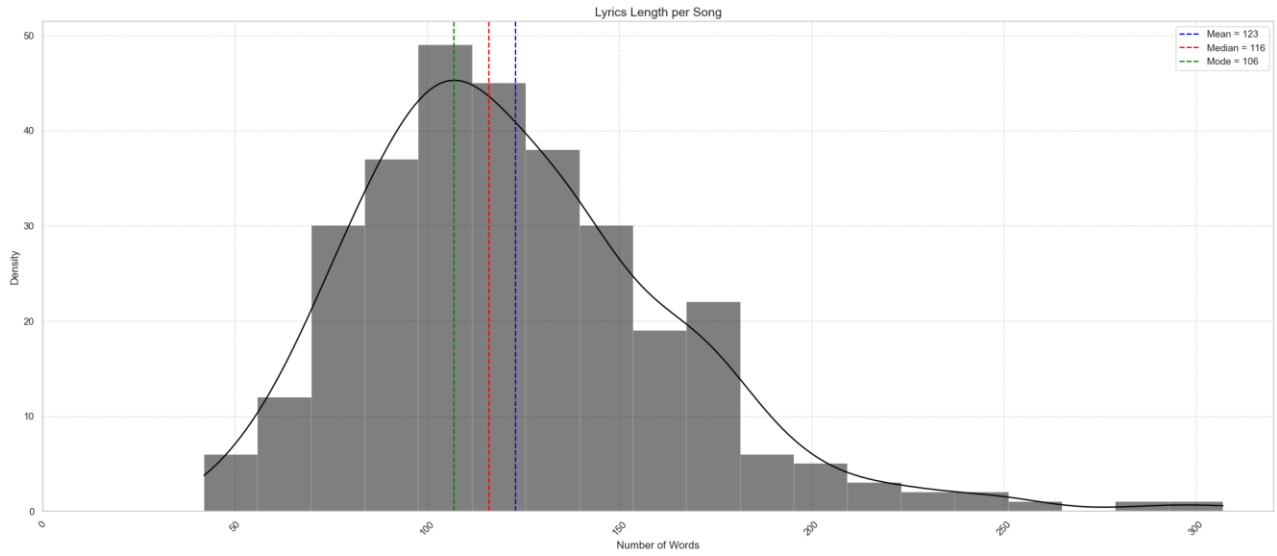

Fig. 4.4.1.1 *Histogram Overlaid with KDE Curve of Lyrics Length per Amr Diab's Song*

As shown in *fig 4.4.1.2*, further statistical fitting reveals that a right-skewed normal distribution fits the data better than a standard normal distribution, highlighting this asymmetry in song lengths.

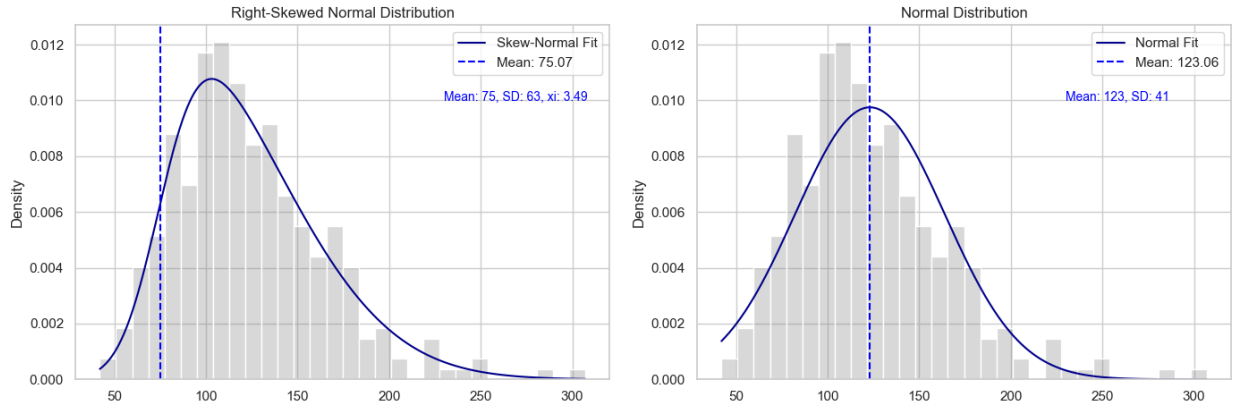

Fig. 4.4.1.2 *Amr Diab's Lyrics Length per Song Against Normal Distribution (Right-Skewed vs. Normal)*

4.4.2 Word Count Per Decade

To investigate whether Amr Diab's songwriting has evolved over time, we analyze the lyrics' word counts per decade. By plotting histograms for each decade (see *fig. 4.4.2*), we examine how the length of his songs has shifted throughout his career.

The dataset reveals some notable trends:

- **1980s:** Early in his career, Amr Diab's songs were longer on average, ranging from 100 to 200 words.
- **1990s:** This decade marks his peak, with a noticeable clustering of song lengths around 150–200 words. Many of his iconic hits emerged during this period, with a steady lyrical style.
- **2000s:** During this period, song lengths began to decrease, indicating a shift towards more concise, commercially viable songs.
- **2010s & 2020s:** The trend towards shorter songs becomes more pronounced, with many tracks falling under 150 words. This shift aligns with the global trend of shorter songs being more successful on streaming platforms.

Interestingly, the longest song in the dataset appeared in the early 2000s, while the shortest song is from the 2020s. This suggests that Amr Diab adapted his songwriting style in response to changing musical tastes and the rise of streaming platforms, where shorter songs tend to perform better.

The analysis of song length by decade shows how Amr Diab adjusted his lyrical approach over time, starting with longer, more detailed compositions and evolving towards shorter, more concise songs suited to modern music consumption habits.

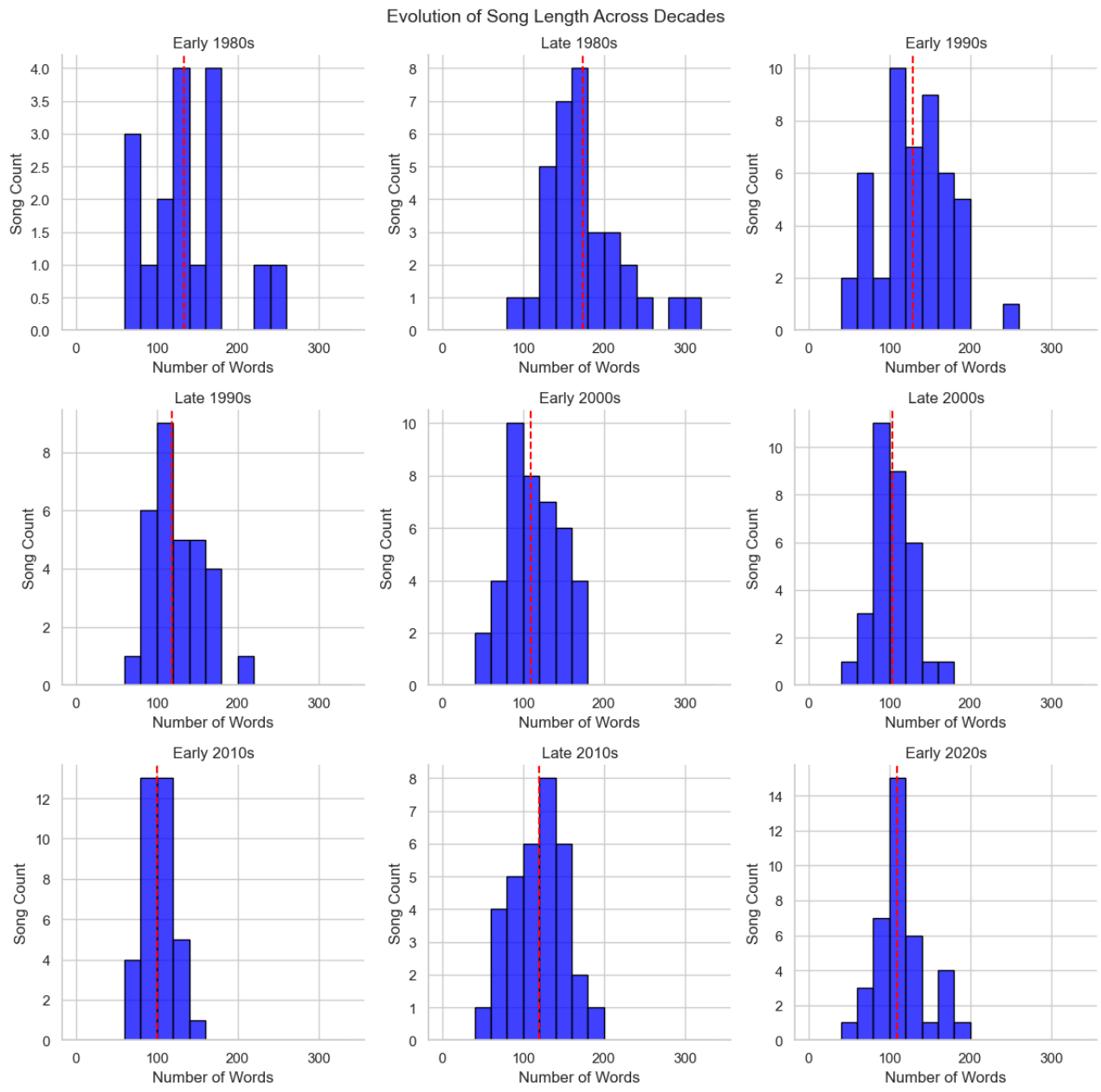

Fig. 4.4.2 *Evolution of Lengths of Amr's Diab's Songs over the Decades*

4.4.3 Popular Words

In this section, we evaluate the most frequently used words across the full set of lyrics in Amr Diab's songs. By identifying these popular words, we can gain valuable insights into the themes and messages that dominate his music. This analysis is based on a simple evaluation of word frequencies, where we process the lyrics to identify the words that appear most often.

To begin, we safely parse the lyrics into lists of words using Python, ensuring that no errors occur if the data format differs from the expected structure. Once the lyrics are converted into word lists, we aggregate all the words into a single collection. We then calculate the frequency of each word, sorting them in descending order to highlight the most common ones. The results show that the most popular words in his songs often include terms related to love, emotions, and personal connection, such as "حبيبي" (my love) and "قلبي" (my heart). These words resonate with listeners, reflecting the intimate and passionate nature of his lyrics.

The most frequent words are visualized using a bar plot (see *fig. 4.4.3.1*), which provides a clear and intuitive view of their relative frequencies. The bar chart emphasizes the dominance of words like "حبيبي," aligning with Amr Diab's focus on themes of love and emotional relationships. The prominence of these words in the lyrics demonstrates how his music consistently communicates feelings of affection and closeness. This focus on emotional depth likely contributes to the widespread appeal of his songs, especially among Arabic-speaking audiences.

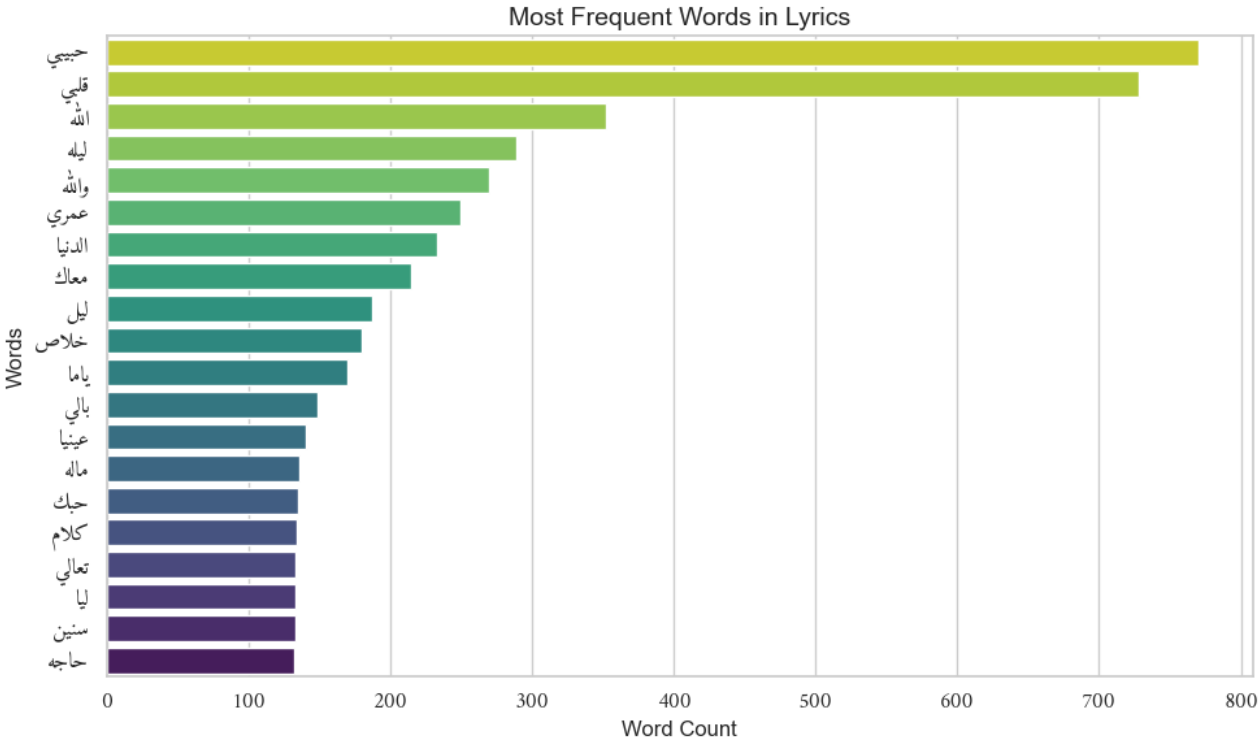

Fig. 4.4.3.1 *Bar Plot of Most Frequent Words in Amr Diab's Lyrics*

In addition to the bar plot, we explore the distribution of word frequencies through a boxplot, as shown in *fig. 4.4.3.2*. By applying a log transformation to the word frequencies, we highlight the spread and outliers in the data. The log-transformed boxplot reveals that while most words cluster within a certain frequency range, a few words, such as "حبيبي," stand out due to their exceptionally

high occurrence. This analysis shows that while most words appear at moderate frequencies, some key terms dominate the lyrics and contribute to the distinctive emotional tone of Amr Diab's music.

These visualizations and statistical analyses provide deeper insights into the lyrical style and thematic focus of Amr Diab's songs. The consistent appearance of certain words reinforces the notion that love, longing, and connection are central to his artistic identity, offering a window into the heart of his musical message.

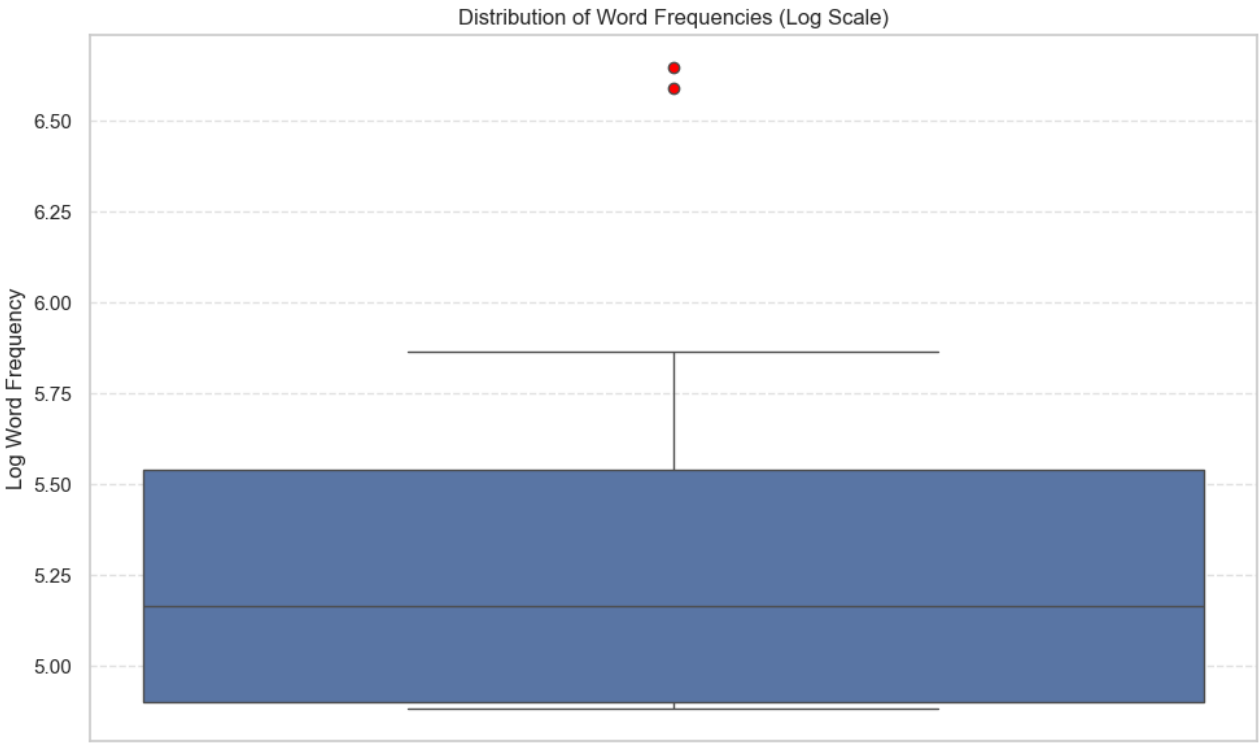

Fig. 4.4.3.2 *Box Plot of Word Frequencies in Amr Diab's Songs*

A word cloud is then used to highlight the most frequently used words in Amr Diab's lyrics, emphasizing those that appear most often (see *fig. 4.4.3.3*). A word cloud is a powerful graphical representation of word frequencies, where the size of each word corresponds to its frequency of occurrence in the text. In this case, the larger words, such as "حبيبي" (my love), stand out clearly, reinforcing their importance and centrality in his music.

Word clouds are not only visually appealing but also incredibly insightful. They serve as an effective tool for summarizing large amounts of text in a simple and digestible format. By presenting words in varying sizes based on their frequency, the word cloud quickly communicates which terms dominate the lyrics. In the context of Amr Diab's songs, the prominence of words related to love, such as "حبيبي" and "قلبي" (my heart), confirms the emotional and romantic themes that run through

his music. This visualization method effectively mirrors our previous findings, where these key terms were identified as central to the emotional depth and connection in his work.

Fig. 4.4.3.3 *Word Cloud of Common Lyrics Words in Amr Diab's Songs*

4.4.4 Popular Words Per Decade

Breaking down the most common words by decade offers deeper insight into how his lyrical content evolved over time. By identifying the top 10 most frequent words per decade, we can trace patterns and shifts in language that reflect not only changes in his personal style but also broader cultural, emotional, and societal influences. For example, certain words might appear more frequently during a particular decade due to the political climate, social mood, or musical trends of that era.

The analysis revealed that some words consistently remained popular across decades, likely reflecting enduring themes in Amr Diab's music, such as love, longing, and emotional expression. However, other terms appeared to be more time-specific, suggesting transient trends or changing focuses in his songwriting. This progression paints a nuanced picture of how his lyrics adapted to different periods while maintaining a core emotional identity. Visualizing this data using horizontal bar charts (see *fig. 4.4.4*), reshaped properly for Arabic text rendering, further highlights these decade-specific differences and reinforces the lyrical shifts in his musical journey.

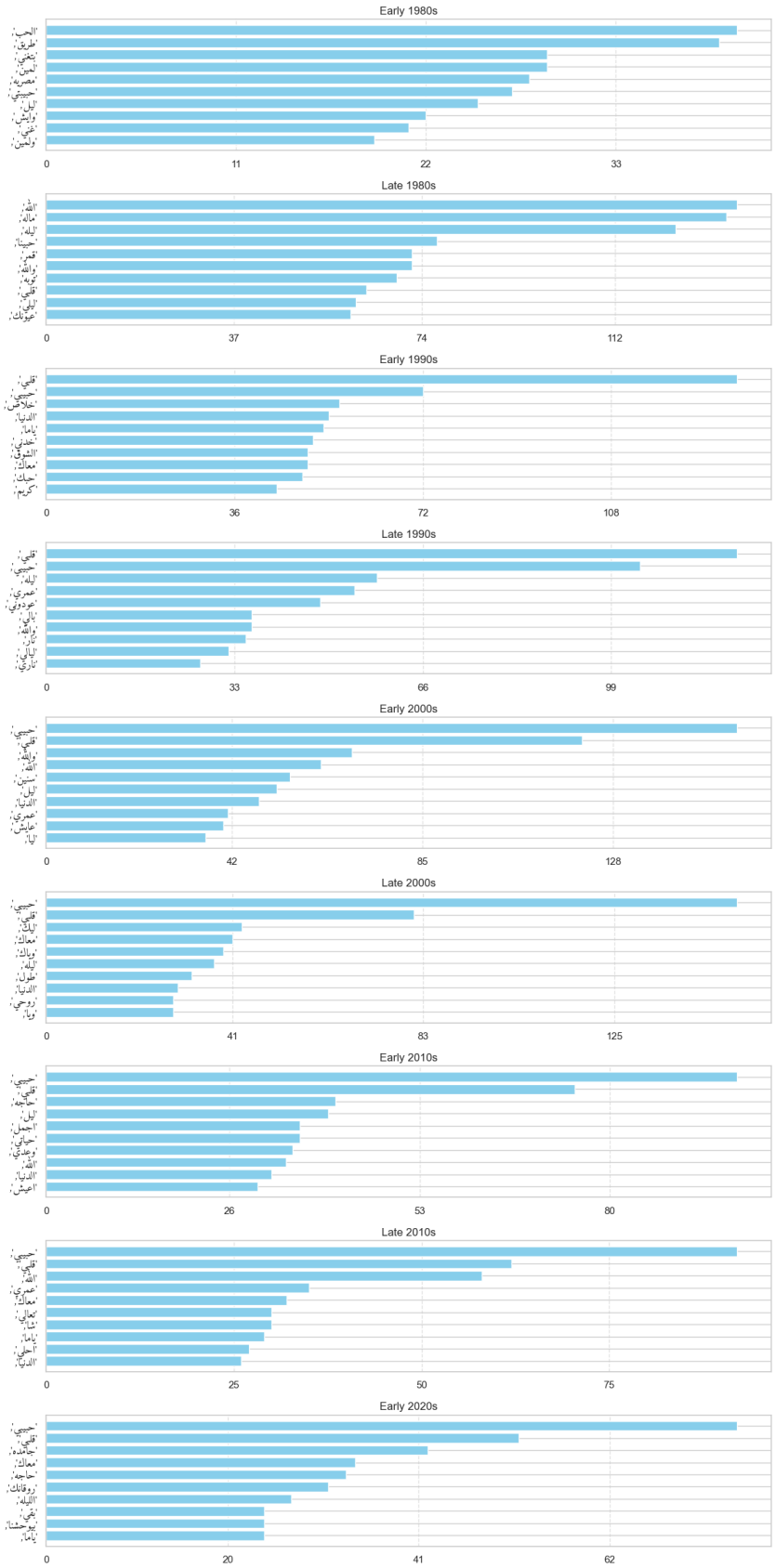

Fig. 4.4.4 *Bar Charts of Top 10 Most Frequent Words in Amr Diab's Lyrics per Decade*

4.4.5 Word Length

Word length is a subtle yet significant feature in songwriting, especially for lyricists striving to match words to rhythm and rhyme. Longer words can be more expressive but are harder to fit into melodic patterns or rhyme schemes, while shorter words enhance fluidity and accessibility. In our analysis of Amr Diab's lyrics, we calculated the length of every word across all his songs and plotted the distribution, overlaid with a Gaussian (normal) curve to capture the overall pattern (see *fig. 4.4.5.1*).

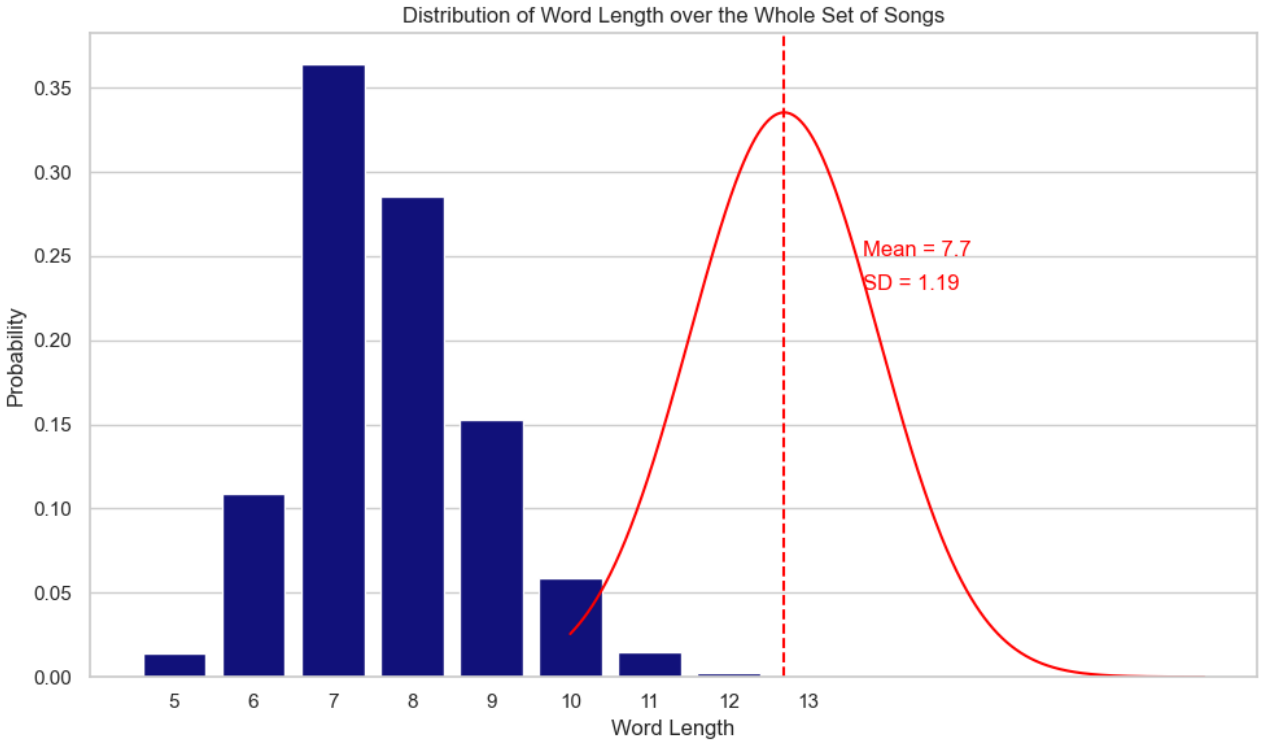

Fig. 4.4.5.1 *Histogram Overlaid with a Gaussian Curve for Word Lengths of Amr Diab's Songs*

The resulting histogram revealed a clear concentration of word lengths around a **mean of 7.7 characters** with a **standard deviation of 1.19**, indicating that most words in his lyrics are of medium length. This aligns well with the expectations of modern pop music, which often emphasizes rhythmic and lyrical flow. The relatively narrow spread suggests a consistent use of words that are neither too long nor too short, helping maintain lyrical cohesion while ensuring the songs remain easy to follow and sing along with.

This preference for mid-length words may reflect Amr Diab's artistic balance between poetic depth and mass appeal. By avoiding overly long or overly simplistic vocabulary, he can deliver meaningful content without sacrificing musicality. Additionally, this may reflect broader trends in contemporary Arabic pop, where clarity, accessibility, and alignment with modern production styles are increasingly prioritized.

To complement our quantitative analysis, we visualized three word clouds highlighting the longest, shortest, and most typical word lengths found in Amr Diab's lyrics (see *figs. 4.4.5.2, 4.4.5.3, 4.4.5.4*). These word clouds provide a visual intuition of how word length correlates with lyrical content and stylistic choices.

Fig. 4.4.5.2 *World Cloud for Longest Words in Amr Diab's Songs*

Fig. 4.4.5.3 *World Cloud for Shortest Words in Amr Diab's Songs*

Fig. 4.4.5.3 *World Cloud for Most Typical Word Lengths in Amr Diab's Songs*

4.4.6 Word Length Per Decade

To understand how Amr Diab's lyrical style has evolved over time, we analyzed the average word length per song across different decades. This gives insight into whether his choice of words, particularly their complexity and length, has changed as his music career progressed from the 1980s to the 2020s. For each song, we calculated the mean word length by dividing the total number of characters in all words by the number of words. Then, we grouped these results by decade to explore how this metric varies over time.

The resulting boxplot visualization (see *fig. 4.4.6.1)* shows that average word length per song has remained remarkably stable over the decades. The median word length generally hovers between 4.5 and 5 characters, with minor fluctuations. For example, songs from the early 1980s and late 1990s show slightly longer average word lengths, while the early 2000s and early 2020s contain songs with shorter averages and more variability. Despite these anomalies and a few notable outliers, the overall pattern suggests that Amr Diab maintained a consistent lyrical style in terms of word complexity, favoring mid-length words throughout his discography. This steadiness might be tied to the accessibility and emotional clarity typical of pop music lyrics.

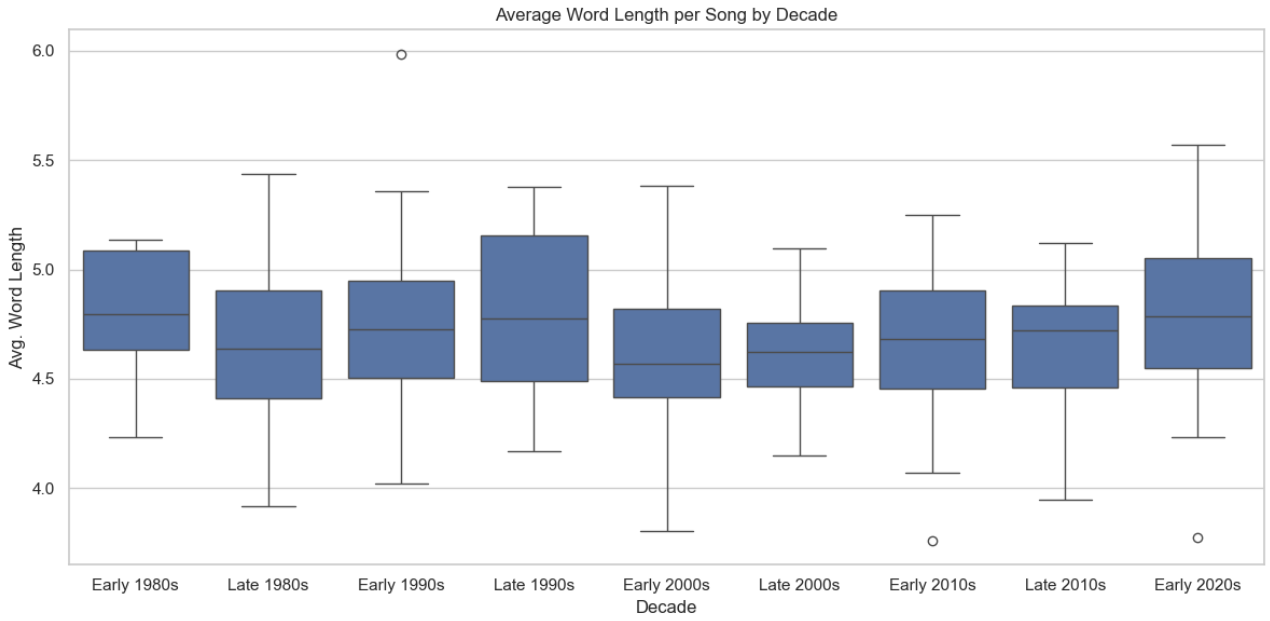

Fig. 4.4.6.1 *Box Plot for the Average Word Length per Amr Diab's Song by Decade*

To delve deeper, we plotted the Probability Density Function (PDF) of average word lengths per song, broken down by decade, as shown in *fig. 4.4.6.2*. Each curve in the plot represents a bell-shaped Gaussian distribution fitted to the distribution of average word lengths for songs within that decade. These curves help visualize not only the average (mean) word length but also the variability (standard deviation) within each era.

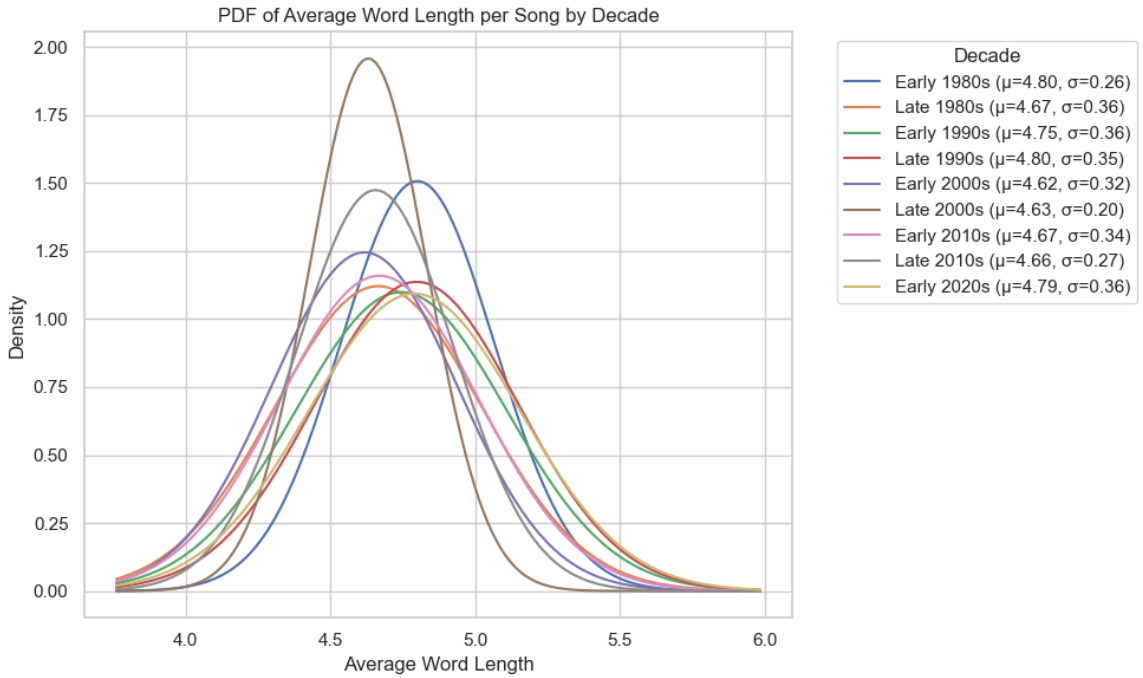

Fig. 4.4.6.2 *PDF of Average World Length per Amr Diab's Song by Decade*

The graph reveals subtle but informative differences. Most decades share a similar central tendency (mean), reinforcing the earlier conclusion of overall lyrical consistency. However, the spread (standard deviation) varies slightly, some decades show tighter curves indicating more uniformity, while others have flatter curves that suggest greater diversity in word length. For instance, the early 2000s exhibit a broader curve, possibly reflecting experimentation in lyrical structure, while the 1990s appear more centered and compact.

Taken together, both the boxplot and the PDF plot demonstrate that although Amr Diab's music has evolved in many stylistic and production-related ways, his choice of word length, a fundamental component of lyrical construction, has remained relatively stable. This consistency may be one of the reasons his music continues to resonate across generations, balancing clarity and expressiveness within a pop framework.

4.4.7 Lexical Diversity

Lexical diversity is a linguistic measure that reflects the range of unique words used within a text. In the context of song lyrics, it quantifies how varied an artist’s vocabulary is within a single song or across their discography. High lexical diversity often indicates a rich and varied vocabulary, while low lexical diversity suggests repetitive or simpler word choices. We analyze the lexical diversity of Amr Diab’s songs to examine whether his lyrical complexity has changed over the decades.

To calculate lexical diversity, we first processed the lyrics by splitting them into words and calculating the number of unique words per song. Songs with missing year or decade information were excluded to ensure accuracy. The analysis was performed by grouping songs by year and calculating the average number of unique words for each group. The data was visualized using a scatter plot to display the diversity trend over time, accompanied by a smooth line to highlight any emerging patterns.

The scatter plot in *figure 4.4.7.1* reveals that the lexical diversity of Amr Diab's songs has remained relatively stable over the years, with most songs exhibiting a diversity score between 30 and 50. However, a few outliers in recent years (2015 to 2020) show significantly higher diversity, suggesting a shift toward more complex or varied lyricism in some of his newer releases.

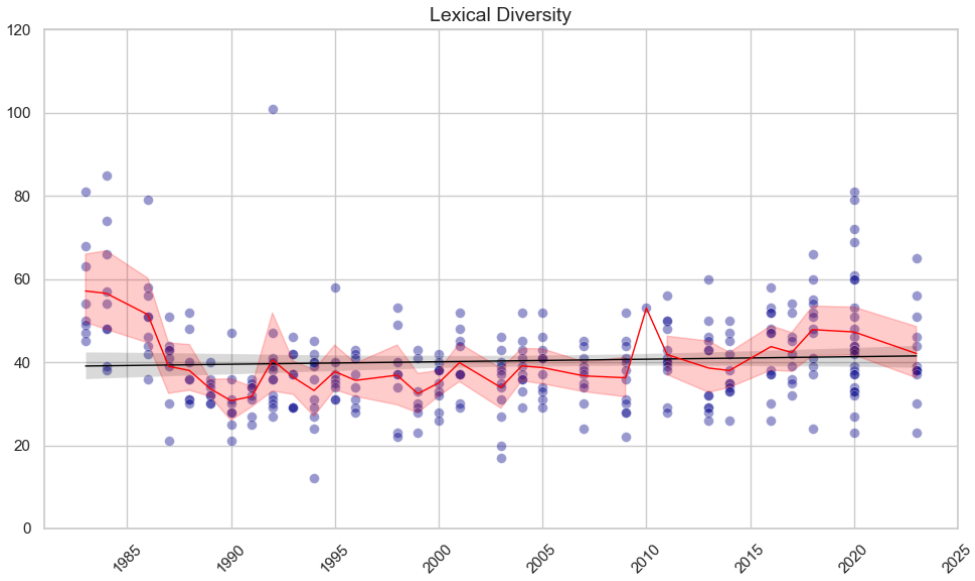

Fig. 4.4.7.1 *Lexical Diversity for Amr Diab's Songs Over the Years*

A trend line superimposed on the scatter plot indicates a slight upward trend in lexical diversity. This could imply a gradual increase in the complexity of Amr Diab's lyrics over time, potentially reflecting evolving songwriting styles or influences.

For a smoother visualization, we created a smoother version of the plot (shown in *fig. 4.4.7.2*) by calculating the average lexical diversity per year. This approach summarized the diversity across all songs released in the same year, providing a single representative point per year. The resulting plot showed a more continuous trend, highlighting the gradual increase in lexical diversity more clearly.

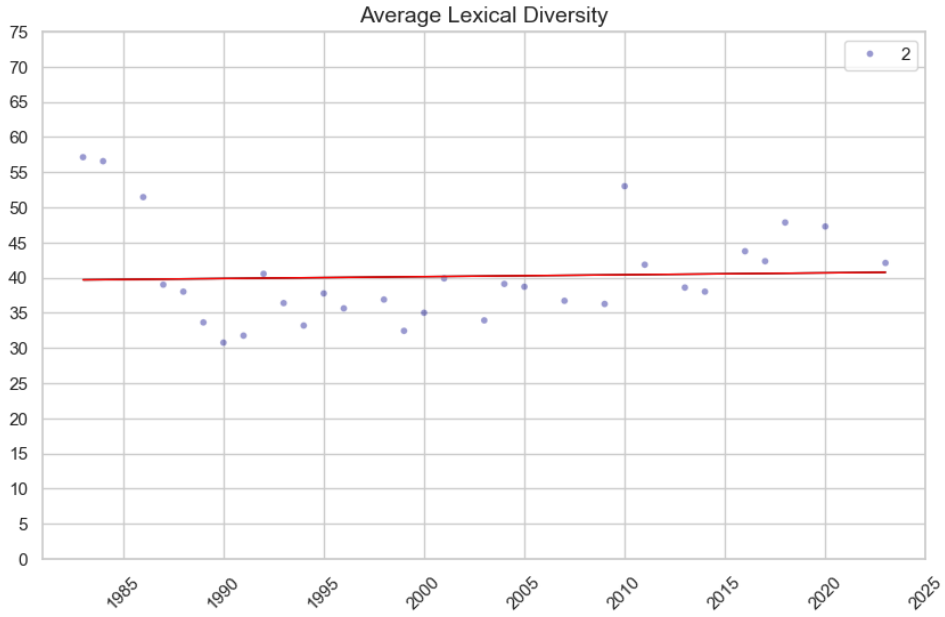

Fig. 4.4.7.2 *Average Lexical Diversity for Amr Diab's Songs Over the Years*

4.4.8 Lexical Density

While the initial analysis focused on absolute measures of lexical diversity, a more nuanced picture can be obtained through relativistic metrics, such as lexical density. Lexical density is defined as the number of unique words in a song divided by the total number of words in that song. It serves as an indicator of word repetition, which is a crucial tool in songwriting. Higher lexical density indicates less repetition, while lower density suggests a more repetitive lyrical structure. It is essential to note that lexical density does not account for sequential repetition, which is another stylistic choice often used by lyricists.

To calculate lexical density, we processed the lyrics by joining all words per song and computing the ratio of unique words to the total word count. Songs missing the necessary information were excluded. We then calculated the average lexical density per year and visualized the results through scatter plots, complemented by linear and loess smoothers to capture trends over time. An additional horizontal line marked the overall average density across all songs.

The scatter plot (see *fig. 4.4.8.1*) indicates that most of Amr Diab's songs have a moderate to high lexical density, highlighting the richness of his lyrical content. However, a slight decline in density in the years between 1990 and 2005 was observed. The decline might be partly attributed to the reduction in the number of songs released. The dashed line represents the overall average lexical density across all songs, calculated to be approximately 40.30.

Fig. 4.4.8.1 *Lexical Density of Amr Diab's Songs Over The Years*

To further clarify the trend, we aggregated songs released in the same year and plotted the average lexical density per year (as shown in *fig. 4.4.8.2*). The resulting graph showed a smoother distribution, with a noticeable decline around 1999 before rising again toward 2020. The highest average lexical density was observed before 1985, followed by a drop and a subsequent gradual increase.

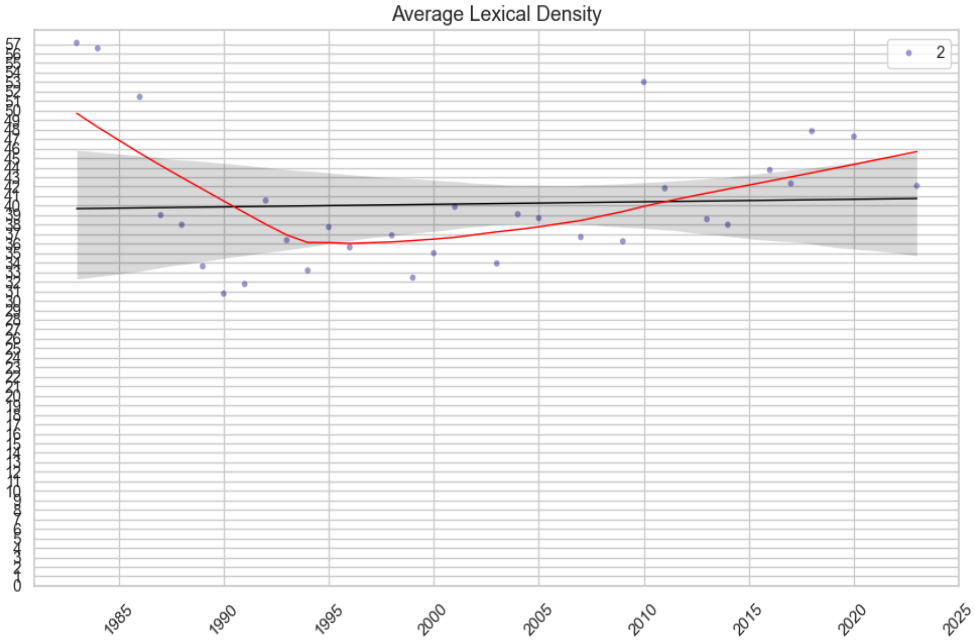

Fig. 4.4.8.2 *Average Lexical Density of Amr Diab's Songs Over the Years*

This pattern may suggest a balance between lyrical length and word diversity, reflecting Amr Diab's evolving songwriting style and the diversity in his musical output across various media, including albums, movies, and concerts.

4.4.9 Lexical Diversity and Density Over Decades

To gain a more comprehensive understanding of Amr Diab's lyrical evolution, we examined both lexical diversity and lexical density on a decade-by-decade basis. We computed the average lexical diversity and lexical density for each decade by grouping the songs according to their release period. To better visualize the changes over time, we reshaped the data and ordered the decades chronologically.

The combined bar plot (in *fig. 4.4.9.1*) highlights an interesting pattern: the lexical diversity and density were relatively high in the early 1980s but declined during the late 1980s and early 1990s. This period likely reflects a phase where Amr Diab's songwriting style was more repetitive or less varied. However, from the early 2000s onward, both metrics began to rise, peaking in the 2010s and stabilizing into the 2020s.

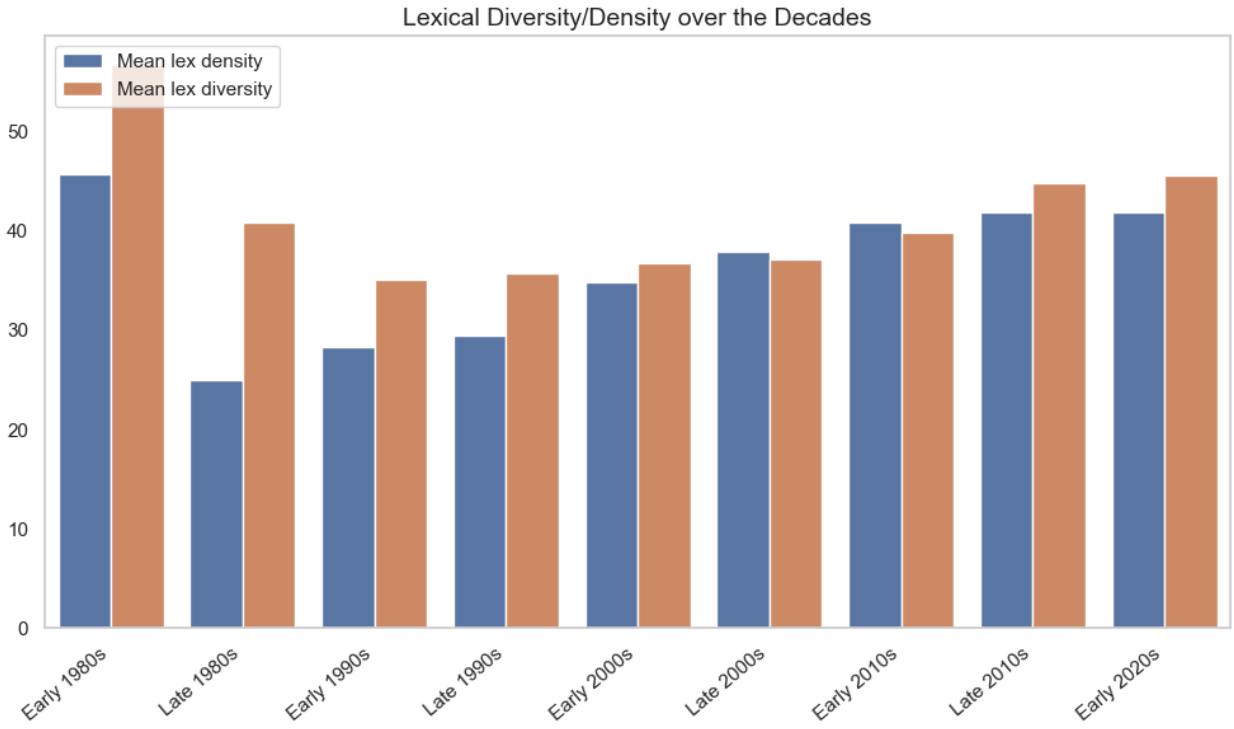

Fig. 4.4.9.1 *Lexical Diversity/Density of Amr Diab's Songs Over the Decades*

One possible interpretation is that the lyrical content became more sophisticated as Amr Diab matured as an artist, possibly influenced by changing musical trends or the demand for more diverse and meaningful lyrics. The increase in lexical density over recent decades might also reflect a strategic shift toward more compact and meaningful lyricism, balancing repetition with novelty.

Overall, the co-evolution of lexical diversity and density indicates a trajectory towards more varied and less repetitive lyrics over time, with some fluctuations that could correspond to specific stylistic or thematic phases in Amr Diab's career.

4.4.10 TF-IDF Analysis

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used to evaluate the importance of a word within a document relative to a collection of documents (corpus). In the context of song lyrics, it helps identify the most distinctive and significant words used by Amr Diab across his discography.

To calculate TF-IDF, we treated each song as a document and the entire discography as a corpus. Term Frequency (TF) measures how often a word appears in a song, while Inverse Document Frequency (IDF) measures how rare a word is across all songs. Multiplying these two values gives

the TF-IDF score, highlighting words that are frequent within a song but uncommon across the entire corpus.

The analysis of TF-IDF scores revealed interesting patterns in Amr Diab's lyrics across different decades. Words related to love, emotions, and personal connections consistently rank among the highest TF-IDF scores, reflecting recurring lyrical themes in his music. For example, words like 'حب' (love), 'قلب' (heart), and 'حبيبي' (my beloved) often appear at the top, indicating their centrality in many songs.

Conversely, the words with the lowest TF-IDF scores tend to be common filler words or frequently repeated terms, such as 'يا' (oh) and 'في' (in). These words are less distinctive because they are used widely across songs and do not significantly differentiate one song from another.

To better understand how lyrical themes have changed over time, we compared the highest and lowest TF-IDF words for each decade. We found that in earlier decades, words related to traditional romance were prominent, such as 'حب' (love), 'قلب' (heart), and 'عين' (eye). In contrast, in later decades, more modern and socially relevant terms started to emerge, like 'حياة' (life), 'زمان' (time), and 'عالم' (world), indicating a thematic shift in Amr Diab's music. Additionally, contemporary songs include words like 'روحي' (my soul) and 'جامدة' (awesome), reflecting evolving language and expressions of emotion.

4.4.11 POS Analysis

Part-of-Speech (POS) analysis, also known as POS tagging, is the process of identifying and labeling each word in a sentence with its corresponding grammatical category, such as noun, verb, adjective, adverb, and so on. This analysis is crucial in understanding the structure and meaning of sentences and plays an important role in various natural language processing (NLP) tasks like machine translation, sentiment analysis, and information extraction. Since Arabic is a morphologically rich language with complex word structures and inflections, accurate POS tagging is particularly challenging.

we utilized CAMeL Tools, a Python toolkit specifically designed for Arabic NLP, to perform POS analysis on Amr Diab's song lyrics. After preprocessing the lyrics, we applied the MLE Disambiguator and the Default POS Tagger from CAMeL Tools to generate POS tags for each word. We then constructed feature vectors representing the frequency of each POS tag per song and compiled a feature matrix to analyze grammatical patterns. To reduce dimensionality and enhance interpretability, Principal Component Analysis (PCA) was applied to the feature matrix. Finally, K-means clustering was employed to identify patterns and group similar songs based on POS features.

The POS analysis revealed that nouns and proper nouns are the most frequent grammatical categories in Amr Diab's lyrics, reflecting a strong thematic focus on personal and romantic connections. Verbs and adjectives also appeared frequently, indicating a dynamic and descriptive quality in the lyrics. The clustering analysis based on POS features showed that songs with higher lexical diversity tend to exhibit a wider range of POS tags, demonstrating a richer grammatical structure.

To determine the optimal number of clusters, we employed both the Elbow method and the Silhouette coefficient. The Elbow method indicated an optimal cluster size of K=8, focusing on minimizing the within-cluster sum of squares. However, the Silhouette analysis, which evaluates both cohesion and separation, suggested that K=3 offers better-defined clusters. The relatively moderate silhouette score of 0.39 suggests that while some distinctions exist between clusters, the overall similarity between songs — likely driven by recurring romantic and emotional themes — results in overlapping clusters. This outcome indicates that Amr Diab's lyrics share a cohesive thematic structure, despite minor stylistic variations.

The graph depicting the Elbow and Silhouette analyses confirms this interpretation, as it demonstrates that the increase in the number of clusters beyond three does not significantly enhance cluster quality. This further supports the choice of K=3 as the most meaningful grouping, highlighting the balanced relationship between cluster compactness and separation.

5 Lyrics Analysis (using LLM's)

In recent years, Large Language Models (LLMs) have emerged as powerful tools in the field of natural language processing (NLP). These models, trained on vast corpora of text data, possess the ability to understand, generate, and manipulate human language with remarkable fluency and contextual accuracy. In our work, we utilize LLMs to analyze lyrics of Amr Diab.

The goal of this section of the project is to explore how advanced LLMs can be employed to detect sentiment and emotion in these lyrics, offering a data-driven lens into the emotional landscape of his musical works. Specifically, we leverage Google's Gemini model, known for its multilingual capabilities and contextual awareness, to perform sentiment classification and emotion detection on a curated dataset of Amr Diab's songs.

5.1 Model Setup

To integrate LLMs into our analysis pipeline, we employed the `google.generativeai` Python library and initialized the Gemini-2.0-Flash model. This particular variant of the Gemini series was chosen for its speed and lightweight performance, making it suitable for batch inference on large datasets. We configured the API with a secure key and confirmed the model's readiness to perform real-time sentiment and emotion analysis on Arabic text.

5.2 Sentiment Analysis

Sentiment analysis aims to classify the general attitude or emotion conveyed in a text, commonly into categories such as positive, negative, or neutral. In this project, we used Gemini to analyze the sentiment of each song in our dataset. Our custom function sent a prompt containing the song's lyrics to the model, requesting an assessment of sentiment along with a concise explanation. Gemini's

understanding of both modern and classical Arabic syntax allowed it to capture nuances such as romantic overtones, poetic imagery, and cultural idioms, which are prevalent in Arabic music.

To illustrate the process, we tested the sentiment analysis function on a randomly selected song. The model not only identified the sentiment as positive, but also provided a thorough breakdown highlighting expressions of love, joy, and commitment. Encouraged by this level of depth, we scaled the process to the entire dataset. We implemented a batching mechanism and error handling logic to manage API rate limits and maintain stability during the analysis phase.

The sentiment results, including both the label and explanatory text, were stored in a new CSV file, enabling further statistical analysis and visualization. We then merged these results with metadata (such as decade of release) and plotted the sentiment distribution across time (see *fig. 5.2*).

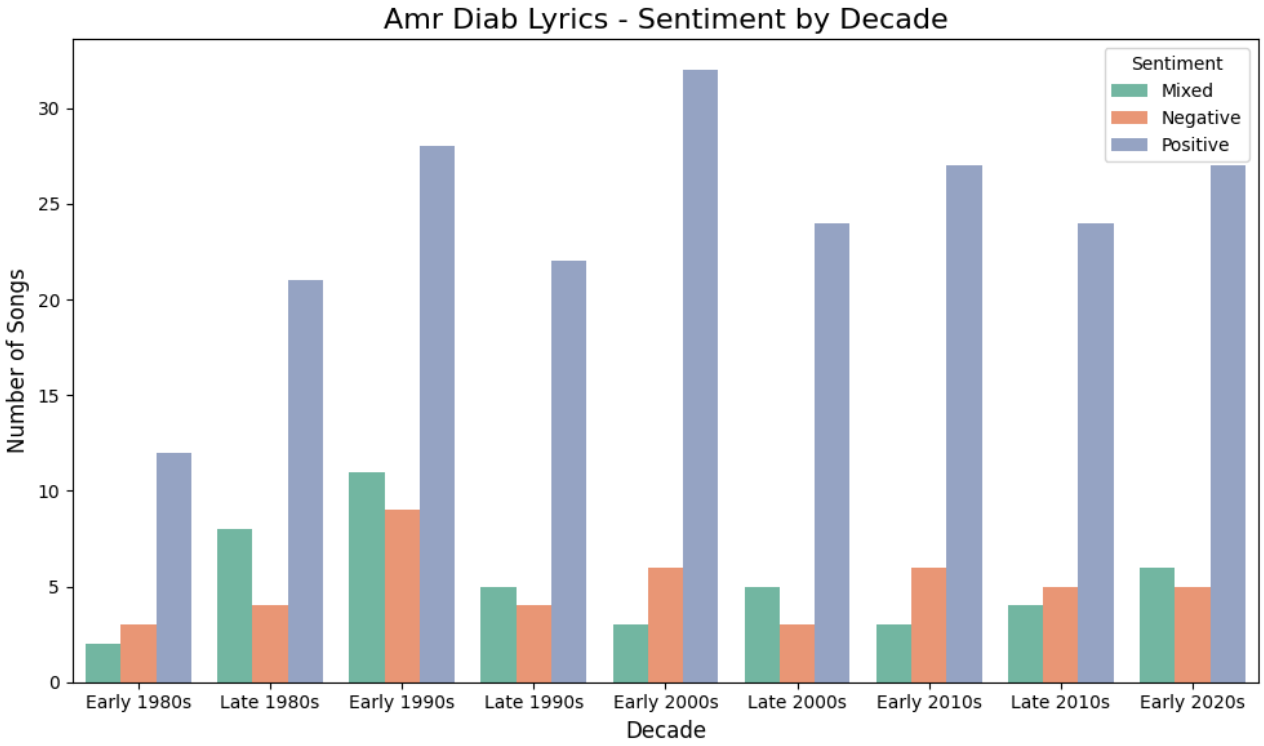

Fig. 5.2 *Sentiment Distribution of Amr Diab's Songs Across Decades*

Our sentiment analysis revealed a dominant trend: Amr Diab's music consistently leans toward positive sentiment. Negative and mixed sentiment songs were significantly fewer, suggesting a stylistic preference for uplifting themes. This trend aligns with the artist's public image and the cultural demand for music that celebrates love, optimism, and emotional connection. Furthermore, we observed that this trend persisted across decades, emphasizing the consistency of Amr Diab's artistic brand and audience appeal.

5.3 Emotion Detection

While sentiment analysis provides a general view of emotional polarity, emotion detection seeks to identify specific emotional states such as love, joy, sadness, or nostalgia. These finer-grained insights are particularly important in lyrical analysis, where emotions are often layered and metaphorically expressed.

In this phase, we utilized Gemini to identify the dominant emotion in each song. A tailored prompt was constructed to instruct the model to return a single, most relevant emotion from a curated list. The model demonstrated strong capability in distinguishing nuanced feelings embedded within poetic and idiomatic Arabic expressions. We validated the model's responses by testing them against randomly selected songs and manually reviewing the coherence of the emotional label with the lyrical content. Results were then plotted as shown in *fig. 5.3*.

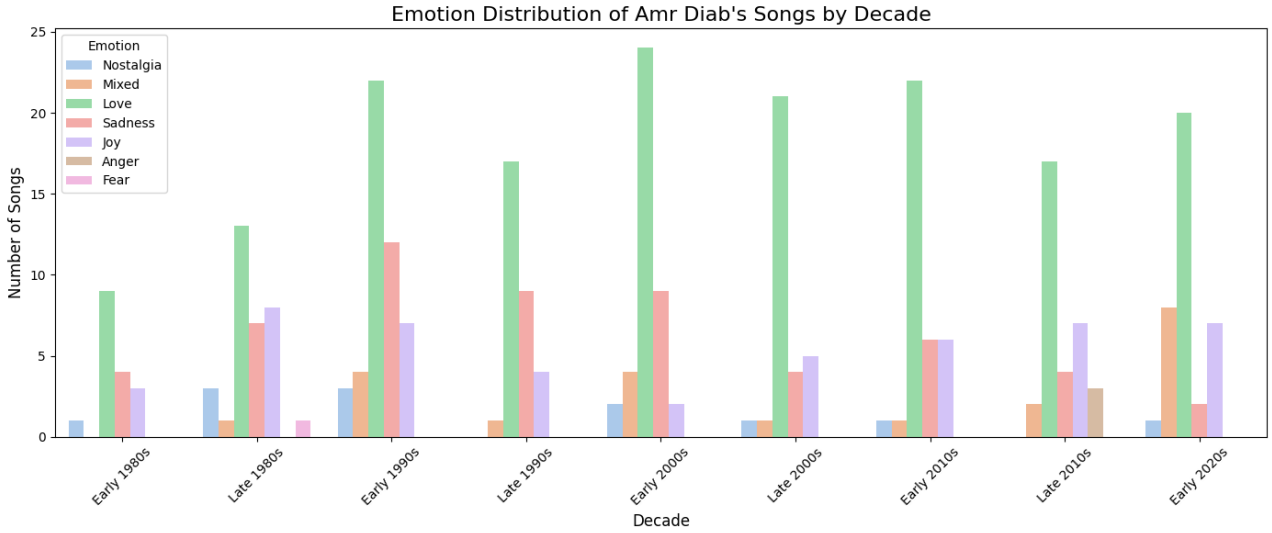

Fig. 5.3 *Emotion Distribution of Amr Diab's Songs Across Decades*

Upon analyzing the emotion distribution across Amr Diab's song lyrics, we observed that the most dominant emotions were joy, sadness, and love. These three emotions consistently appeared across a significant portion of the dataset, reflecting the recurring emotional themes in his musical style. Songs associated with joy often featured upbeat rhythms and optimistic expressions of love or personal freedom. In contrast, sadness-driven lyrics were typically slower and dealt with themes of longing, heartbreak, or solitude.

The emotion love emerged as a central theme, often overlapping with both joy and sadness, depending on the context of the lyrics. This suggests that Amr Diab's artistic expression frequently revolves around romantic experiences, both uplifting and melancholic. Lesser-represented emotions such as anger and fear appeared sporadically, usually in specific narrative-driven songs or dramatic verses. Overall, the emotional trends highlight Amr Diab's versatility in expressing a wide emotional range, while maintaining a consistent emphasis on themes of love and personal sentiment.

5.4 Topic Modeling

In our analysis of Amr Diab's lyrical corpus, we extended our exploration beyond sentiment and emotion by applying topic modeling, also known as theme detection. While sentiment analysis concerns itself with the overall tone of the lyrics, positive, negative, or neutral, and emotion detection aims to identify specific feelings such as joy, sadness, or anger, topic modeling addresses a fundamentally different question: *What is the song about?* Rather than evaluating how the artist feels or expresses themselves, theme detection aims to uncover the main subject or idea communicated through the lyrics. Examples of such themes include love, heartbreak, personal growth, friendships, farewells, or even societal issues.

Like in sentiment analysis and emotion detection, we implemented topic modeling by designing a function that leverages a large language model (LLM) to infer the primary theme from the lyrics of each song. The model was prompted with instructions to select from a predefined set of common lyrical themes, such as romantic relationships, goodbyes, social issues, new in love, and so on, and to provide both a theme label and a brief explanatory justification. Results were then plotted as shown in *fig. 5.4*.

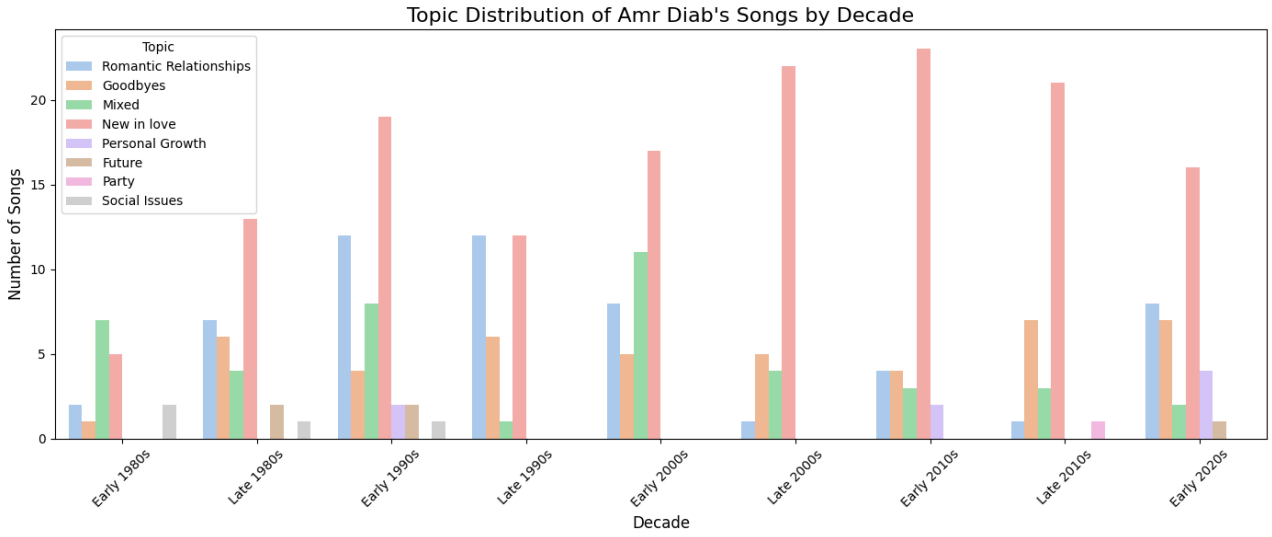

Fig. 5.3 *Topic Distribution of Amr Diab's Songs Across Decades*

Our findings show a clear thematic preference in Amr Diab's discography. The most common theme was "New in love", reflecting the excitement and emotional richness that often accompanies the start of a romantic relationship. This was followed by "Romantic Relationships" , representing ongoing or more established love stories. The third most prevalent theme was "Goodbyes", signaling a recurring focus on loss and separation, a narrative thread that resonates deeply with listeners navigating breakups or emotional farewells.

Interestingly, themes that fall outside the romantic domain were far less represented. For instance, only very few songs addressed social issues, or explored personal growth, and a mere handful dealt with family, friendships, or the future. This indicates that while Amr Diab occasionally touches on broader life themes, his lyrical universe remains heavily centered on romantic and emotional storytelling.

5.5 Analysis of Interrelations Between Emotion, Sentiment, and Theme

In this section, we explore the intricate relationships between emotion, sentiment, and thematic elements in Amr Diab's lyrical corpus. By integrating predictions from emotion classification, sentiment analysis, and topic modeling, we aim to uncover patterns that characterize the emotional and thematic landscape of his music.

5.5.1 Relationship Between Emotion and Sentiment

To examine how emotional tone aligns with sentiment polarity, we began by merging the emotion and sentiment datasets on the common identifier, *Song*. This allowed us to associate each song with both an emotion label (e.g., joy, sadness, anger) and a sentiment label (i.e., positive, neutral, negative).

We visualized this relationship using a *countplot*, as shown in *fig. 5.5.1.1*, that displays the distribution of sentiment labels across different emotion categories. We noticed from the resulting plot, that songs categorized with love or joy overwhelmingly correspond to a positive sentiment. This finding is consistent with expectations, as joyful expressions typically evoke upbeat and optimistic connotations. Emotions such as sadness and anger, in contrast, skew more heavily toward negative sentiments, illustrating a predictable emotional polarity.

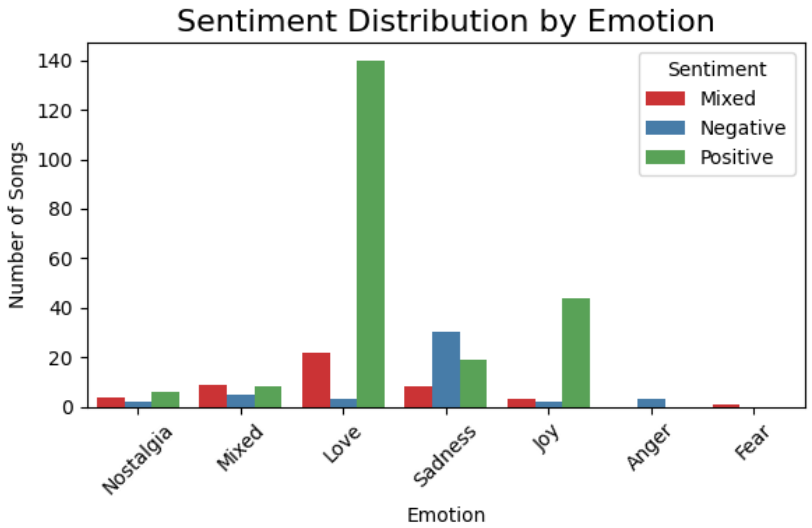

Fig. 5.5.1.1 *Sentiment Distribution by Emotion of Amr Diab's Songs*

To further quantify this association, we generated a *heatmap,* as shown in *fig. 5.5.1.2,* that aggregates song counts by emotion and sentiment labels. The heatmap confirms the dominant presence of positive sentiment in romantic and joyful songs, and negative sentiment in songs evoking sadness and anger. This dual-plot analysis highlights the emotional consistency embedded in sentiment-labeled songs and reinforces the reliability of our classification models.

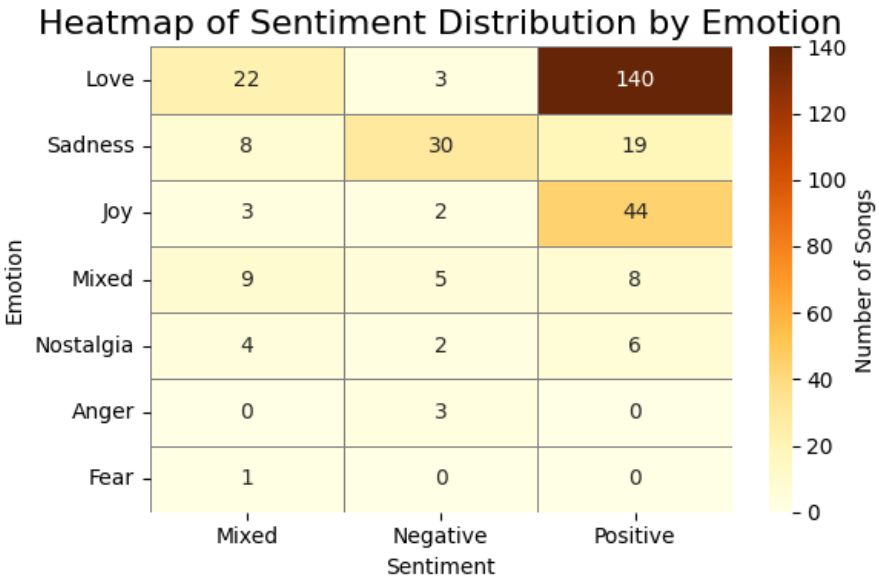

Fig. 5.5.1.2 *Heatmap of Sentiment Distribution by Emotion of Amr Diab's Songs*

5.5.2 Relationship Between Emotion and Theme

Next, we investigated how emotional categories align with lyrical themes. We merged the *emotion_df* and *topic_df* datasets on the *Song* identifier, thereby mapping each song to both an emotion and a thematic label (e.g., New in Love, Romantic Relationships, Goodbyes).

The *countplot* shown in *fig. 5.5.2.1* depicting this relationship revealed that "New in Love" emerges as a dominant theme associated primarily with love and joy. This suggests that Amr Diab's lyrics often celebrate the exhilaration and optimism associated with the beginning stages of romantic relationships. Conversely, the theme "Goodbyes" is most strongly linked with sadness, reflecting emotional narratives centered around loss, heartbreak, or separation. The theme "Romantic Relationships" spans a more diverse emotional spectrum, encompassing joy, trust, and sadness, which points to the multifaceted nature of enduring romantic experiences.

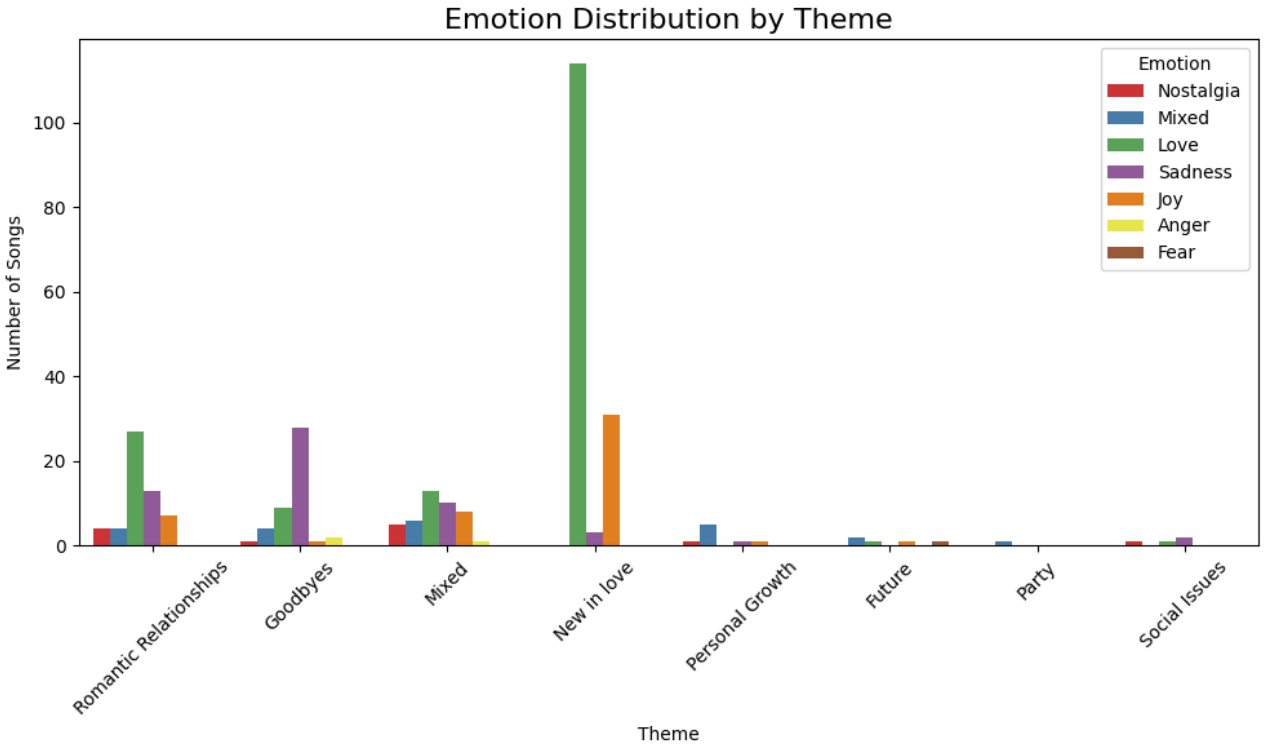

Fig. 5.5.2.1 *Emotion Distribution by Theme of Amr Diab's Songs*

To complement the count-based visualization, we constructed a *heatmap*, shown in *fig. 5.5.2.2*, showing the distribution of emotion labels across themes. This matrix made it evident that themes such as "Friendships", "Personal Growth", and "Social Issues" occur relatively infrequently and exhibit a more balanced spread of emotions. These findings emphasize that while Amr Diab's lyrics are grounded in emotional expression, they are overwhelmingly directed toward romantic and interpersonal contexts, with emotional tones tailored to the narrative of each theme.

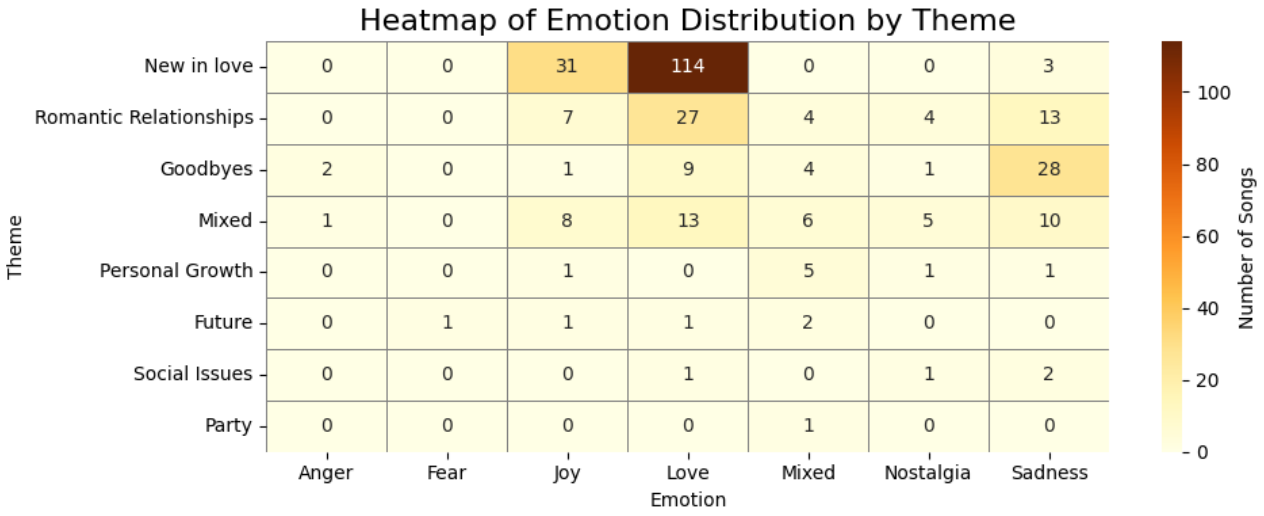

Fig. 5.5.2.2 *Heatmap of Emotion Distribution by Theme of Amr Diab's Songs*

5.5.3 Relationship Between Sentiment and Theme

To complete the triadic analysis, we studied the correlation between sentiment polarity and lyrical theme. We again merged datasets on the *Song* field, this time combining *sentiment_df* and *topic_df*. The resulting plot, shown in *fig. 5.5.3.1,* demonstrated that positive sentiment dominates within the theme "New in Love", reinforcing earlier observations about the celebratory tone of Amr Diab's romantic works. This theme, more than any other, reflects an overwhelmingly optimistic outlook.

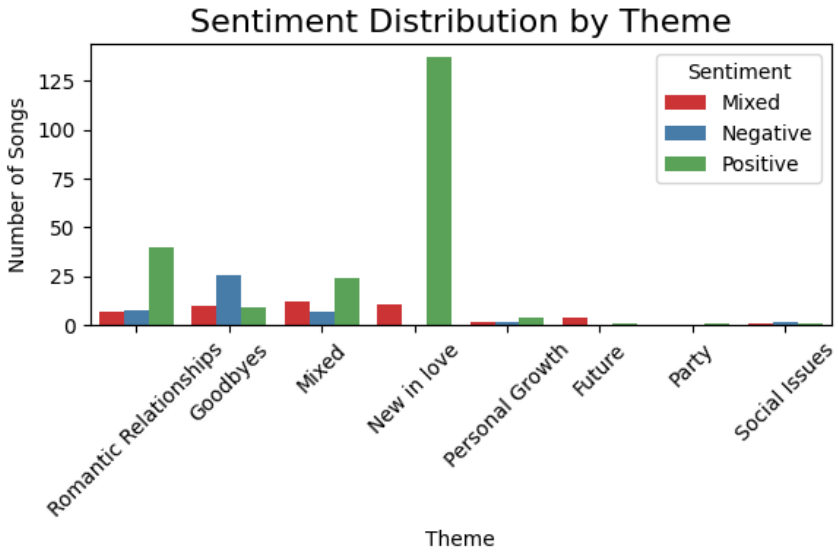

Fig. 5.5.3.1 *Sentiment Distribution by Theme of Amr Diab's Songs*

The theme "Romantic Relationships" presents a more varied sentiment profile. While positive sentiment remains predominant, there is a notable presence of neutral and negative sentiments, suggesting that this category includes both uplifting and emotionally complex songs. In stark contrast, the "Goodbyes" theme is characterized almost entirely by negative sentiment, consistent with themes of departure, longing, and heartbreak.

A sentiment-by-theme *heatmap*, shown in *fig. 5.5.3.2*, further validated these insights, revealing that while certain themes exhibit polarity extremes (e.g., New in Love as mostly positive, Goodbyes as mostly negative), others such as "Social Issues" and "Personal Growth" maintain a more even distribution. These less frequent themes appear to lack a strong emotional or sentimental bias, potentially serving as lyrical interludes to explore broader reflections or observations.

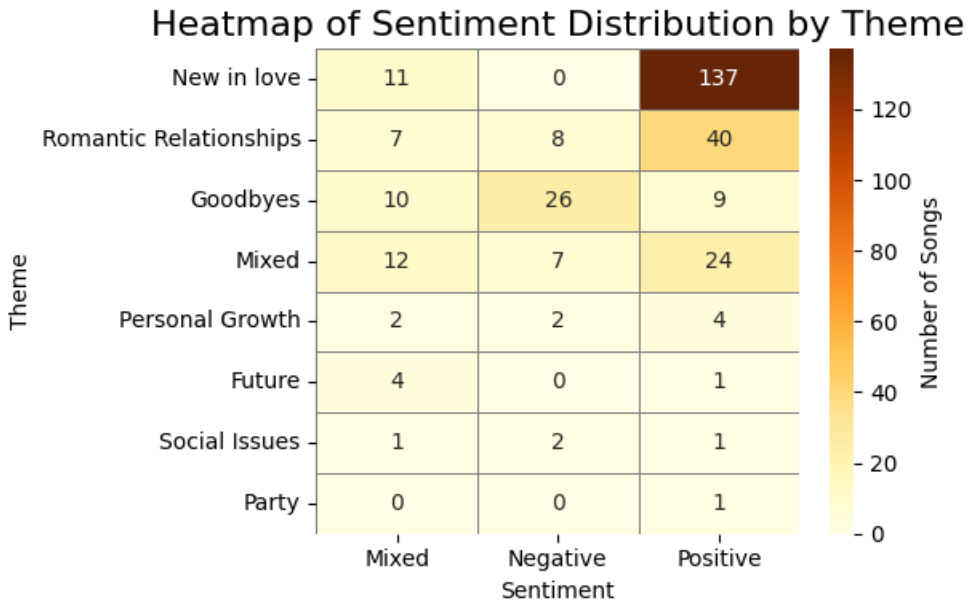

Fig. 5.5.3.2 *Heatmap of Sentiment Distribution by Theme of Amr Diab's Songs*

5.6 Text Embeddings and Clustering

In this section, we explore how semantic embeddings can be utilized to represent the lyrical content of Amr Diab's songs as numerical vectors, thereby enabling meaningful computational analysis. Traditional text analysis techniques such as Bag-of-Words or TF-IDF primarily rely on lexical features like term frequency, which often fail to capture the nuanced meanings or contextual similarities across different texts. To overcome these limitations, we employ text embeddings, a modern technique rooted in deep learning, which translates each song's lyrics into a dense vector of real numbers that encapsulate its semantic structure. These embeddings enable us to compare songs based on their underlying meanings, even if they do not share surface-level vocabulary.

5.6.1 Generating Embeddings

To generate these embeddings, we use the `embedding-001` model provided by the Gemini API. Each song's lyrics are passed to the API with the `semantic_similarity` task type specified, ensuring that the resulting vectors are well-suited for grouping semantically similar texts. The resulting vectors are then stored in a structured format for further analysis.

Once the embeddings are generated and stored, we proceed to perform unsupervised clustering using the KMeans algorithm. Prior to clustering, we apply Principal Component Analysis (PCA) to reduce the dimensionality of the embeddings from hundreds of dimensions down to 50. This step not only enhances computational efficiency but also improves the separation of clusters in subsequent visualization. With the embeddings reduced, we then apply KMeans clustering with five clusters, hypothesizing that the lyrical themes of Amr Diab's songs naturally segment into a small set of distinct groups.

Following the clustering, we assign a cluster label to each song and visualize the results using a 2D scatter plot derived from the top two PCA components, as shown in *fig. 5.6.1.1*. This allows us to inspect the distribution of songs and the separation of clusters visually. The clustering reveals underlying semantic groupings that suggest shared themes, emotions, or narrative arcs in the lyrics.

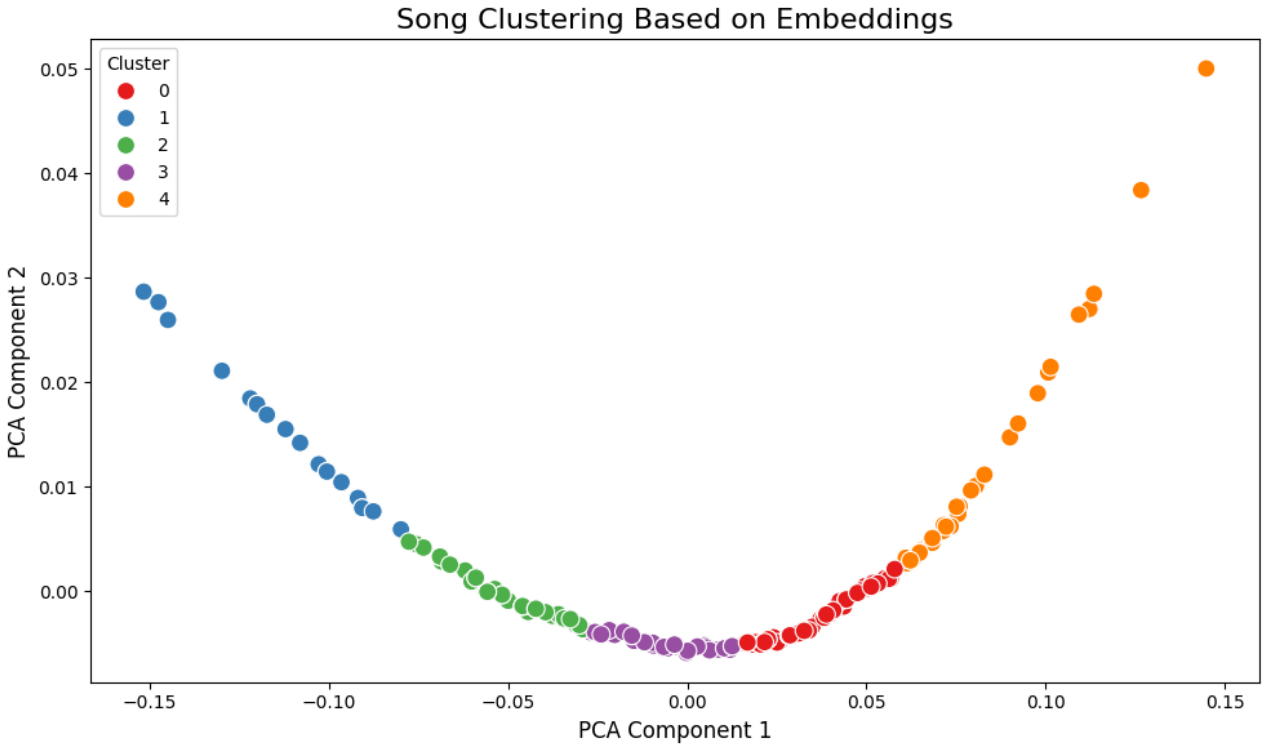

Fig. 5.6.1.1 *Scatter Plot of Song Clusters of Amr Diab's Songs*

To further analyze the nature of these clusters, we merge the clustering results with three previously computed annotations: thematic topics, sentiment labels, and emotion labels, and plot them as shown in *fig. 5.6.1.2, fig. 5.6.1.3,* and *fig. 5.6.1.4.* This enriched dataset allows us to examine how these attributes vary across clusters.

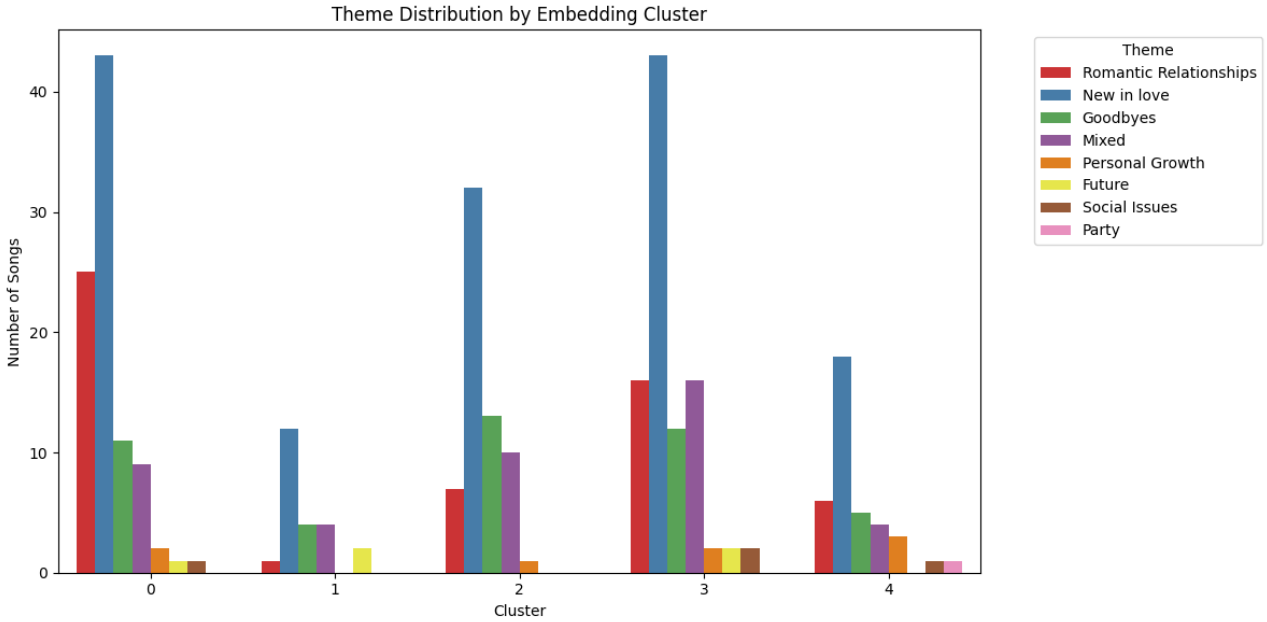

Fig. 5.6.1.2 *Theme Distribution by Embedding Cluster of Amr Diab's Songs*

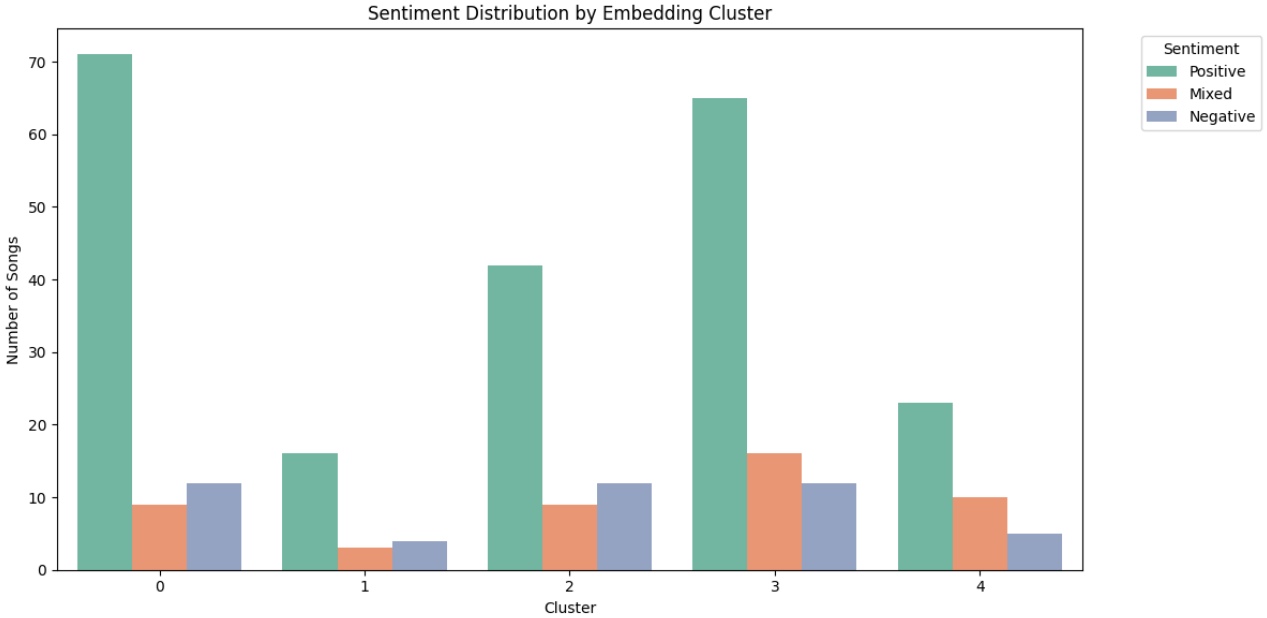

Fig. 5.6.1.3 *Sentiment Distribution by Embedding Cluster of Amr Diab's Songs*

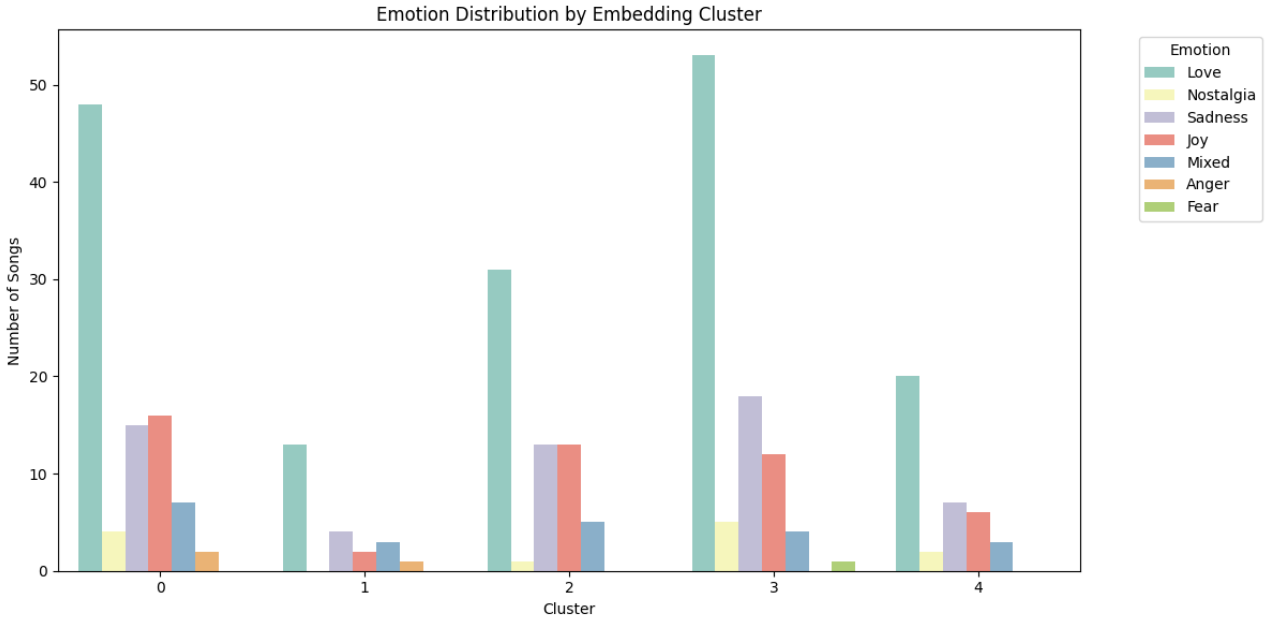

Fig. 5.6.1.4 *Emotion Distribution by Embedding Cluster of Amr Diab's Songs*

5.6.2 Observations from Clustering

Our visualizations indicate that the theme "New in Love" is prevalent across all clusters, underscoring Amr Diab's frequent lyrical focus on romantic beginnings.

Each cluster, however, reveals distinct thematic and emotional profiles. **Cluster 0** is dominated by songs centered around romantic love and new relationships, with high levels of positive sentiment. In contrast, **Cluster 1** shows a broader thematic mix, including *goodbyes* and *future-focused* narratives, and has a more balanced emotional distribution with notable presence of *sadness* and *mixed emotions*. **Cluster 2** shares similarities with Cluster 1 but leans more toward positive sentiment and a blend of *joy* and *sadness*.

Cluster 3 returns to themes of romantic relationships but with greater emotional intensity and a clear tilt toward *love* and *positive emotions*, while **Cluster 4** stands out as the most diverse in its content. It presents a relatively even distribution of themes, sentiments, and emotions, suggesting that this group captures songs with more complex or less conventional lyrical structures.

6 Entropy analysis

To gain deeper insight into the linguistic complexity of Amr Diab's songs, we performed an entropy analysis on the lyrics. Entropy, a concept from information theory, quantifies the unpredictability or

diversity of a text, in this case, based on the distribution of letters. By computing entropy across different scopes, including the entire corpus, individual songs, years, and decades, we aim to assess how varied or repetitive the lyrics are and how this has evolved over time. This analysis helps us better understand trends in Amr Diab's songwriting style, such as whether his lyrics have become more diverse or complex throughout his career.

6.1 Calculating Entropy

To quantify the diversity and unpredictability in the song lyrics, we first defined a function that computes the entropy of a given probability distribution. Additionally, we created a function to compute the empirical probability mass function based on the sample data. A crucial preprocessing step involved normalizing various forms of certain Arabic letters into a single unified form. For example, different forms of the letter "Alif" such as أ and ا were all treated as the same letter to ensure consistent frequency counting.

We then computed the entropy for the entire corpus of lyrics, which was found to be approximately 2.512. This value represents the overall uncertainty or linguistic variability within the entire collection of songs.

Next, we calculated the entropy individually for each song in the corpus. The song with the lowest entropy was identified as **"لا لا"**, with an entropy value of **1.623**, indicating that this song's lyrics are relatively more predictable and less diverse in terms of letter usage. On the other hand, the song with the highest entropy was **"بلاش تكلمها"**, which had an entropy value of **2.868**, reflecting a more varied and less predictable distribution of letters in its lyrics.

Figure 6.1 presents a histogram of entropy values for all individual songs. The distribution closely resembles a Gaussian (normal) distribution centered around 2.5, which suggests that most songs have a moderate level of linguistic complexity. A small number of songs fall at the lower and higher extremes, as reflected by the minimum and maximum entropy values noted above. This distribution highlights that while many songs share a similar level of letter diversity, some are distinctly simpler or more complex in their linguistic makeup.

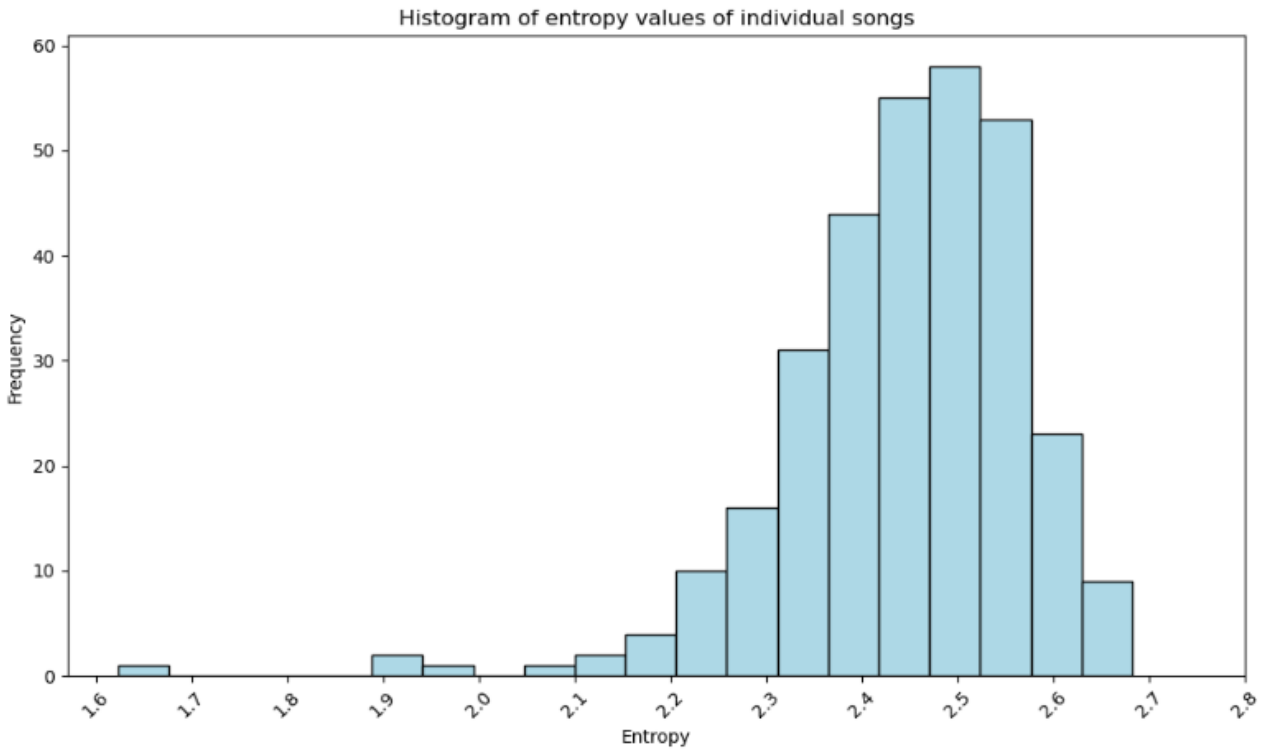

Fig. 6.1 *Histogram of Entropy Values of Individual Amr Diab's Songs*

6.2 Entropy per year

In this section, we analyze the overall entropy of Amr Diab's song lyrics for each year. For any given year, the corpus considered consists of all the lyrics performed during that year. We begin by loading the full songs corpus and filtering out songs without available lyrics to ensure accurate entropy calculations.

Figure 6.2 illustrates the yearly entropy values as individual data points (blue dots). In addition, it features two trend lines: a red LOESS smoothing curve that highlights local variations and a black linear regression line that represents the general trend over time. The plot reveals a modest **increasing trend** in entropy across the years, suggesting that the lyrical complexity and diversity of letter usage in Amr Diab's songs have gradually grown over the span of his career. This increase in entropy can be interpreted as an expansion in the semantic richness or unpredictability of the lyrics over time.

Notably, the LOESS curve suggests a minimum entropy around the early 1980s, followed by a steady rise culminating in a maximum near the early 2020s. This pattern may reflect evolving lyrical styles or themes in Amr Diab's music, indicating a trend toward greater linguistic complexity in more recent works.

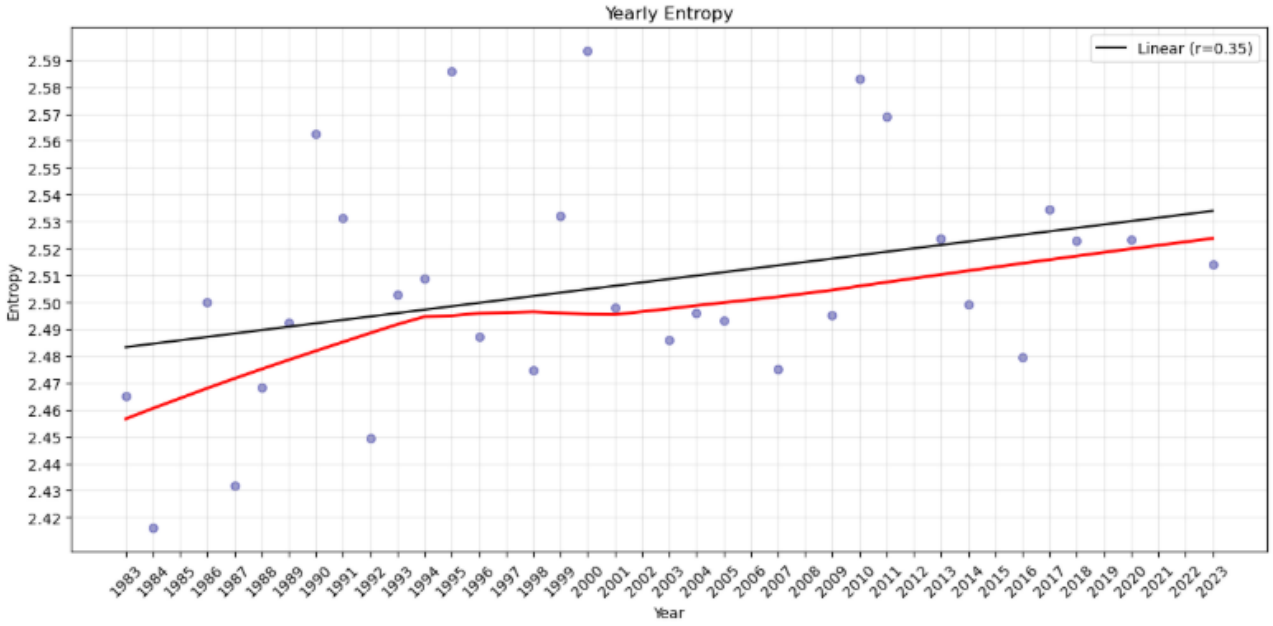

Fig. 6.2 *Yearly Entropy of Amr Diab's Songs*

6.3 Entropy per decade

In this section we compute the overall entropy of all songs performed in each given decade. So, given a particular decade, the corpus over which the entropy is computed is all the lyrics performed during that decade. The corpus over which the entropy is computed is the set of lyrics performed in a given decade. *Figure 6.3* plots the entropies over the decades. As seen in the figure, in all decades the entropy is almost the same for the lyrics performed in that decade. This should then be more accurate, as it computes over a larger text. It is noticed that almost all decades are very close to each other.

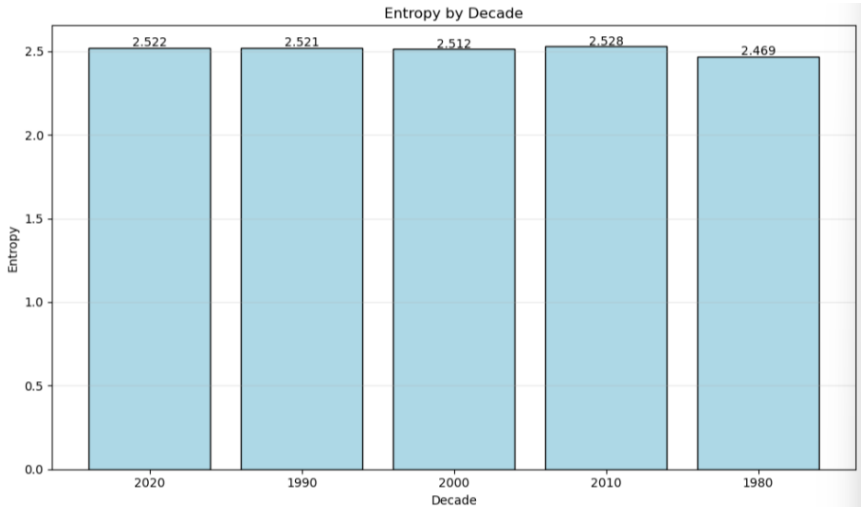

Fig. 6.3 *Entropy by Decade of Amr Diab's Songs*

7 Conclusion

This research paper presents a comprehensive analysis of Amr Diab's lyrical corpus, combining traditional linguistic techniques with modern machine learning methods. Starting from rigorous data preprocessing, including text cleaning, tokenization, and decade categorization, we explored the evolution and patterns of his lyrics over four decades. Through statistical metrics like word frequency, word length, lexical diversity, and entropy, we uncovered how Diab's lyrical style has changed over time, reflecting broader shifts in musical and cultural trends.

Specific word analyses, such as the use of the word "حب" (love), offered insight into recurring themes in his music, while temporal studies of singing rate and word count revealed how his delivery and verbosity evolved with time. Text mining methods further highlighted lexical richness and density, as well as the most frequent and popular words across different decades.

More advanced analyses using large language models (LLMs) provided a deeper semantic layer to our findings. Sentiment analysis and emotion detection captured the affective tone of his lyrics, while topic modeling revealed dominant themes across his discography. By examining relationships between emotion, sentiment, and lyrical themes, and clustering lyrics based on semantic embeddings, we uncovered meaningful groupings of songs that share linguistic or emotional characteristics.

Finally, entropy analysis quantified the unpredictability of Diab's lyrics at both the song and decade levels, showing a gradual increase in linguistic complexity throughout his career.

Together, these analyses paint a rich picture of Amr Diab's musical journey, not only as a legendary performer but also as a lyricist whose language has grown in depth, nuance, and complexity over time. This multifaceted approach sets the stage for future work in Arabic lyrics analysis and helps preserve and appreciate the linguistic artistry of one of the Arab world's most iconic musical figures.