Stylistic analysis of Hebrew song lyrics

Meitar Yeruham & Dean Amar

git repository: https://github.com/kodolic/Stylish-hebrew-songs.git

Goal

The aim of this project is to analyze the stylistic elements of Hebrew song lyrics. The **motivation** stems from a controversial statement made by singer Yehoram Gaon in 2021, who criticized "oriental" (Mizrahi) music for its alleged poor and ungrammatical language. This project seeks to explore stylistic differences across various groups of songs, focusing on aspects such as vocabulary richness, syntactic complexity, and thematic diversity.

Dataset

The project utilizes a large corpus of nearly 15,000 Hebrew song lyrics, downloaded from Kaggle. For each song, additional features such as the performer's year of birth, music style, and song release year were manually associated.

Explanation of Features in the Song Class

1. name

- **Description**: The name of the song.
- **Calculation**: Directly from the dataset.

2. artist

- **Description**: The name of the artist who performed the song.
- **Calculation**: Directly from the dataset.

3. words

- **Description**: Song's lyrics.
- **Calculation**: Directly from the dataset.

4. wordCount

- **Description**: The total number of words in the song's lyrics.
- **Calculation**: Directly from the dataset.

5. uniqueWords

- **Description**: The number of unique words in the song's lyrics.
- **Calculation**: Directly from the dataset.

6. releaseYear

- **Description**: The year the song was released.
- **Calculation**: Retrieves the release year of a song using Spotify's API, Wikipedia, and Shironet, selecting the first available valid year from these sources.

7. songInEnglish

- **Description**: The English translation of the song's lyrics.
- **Calculation**: Translates song lyrics from Hebrew to English using Google Translate and a machine learning translation model, combining chunks of translated text for complete lyrics representation.

8. translatedWords

- **Description**: A list of words from the English translation of the song's lyrics.
- **Calculation**: Splits the English translation of the song's lyrics into individual words using regular expression-based delimiters like periods, commas, and spaces

9. bigrams

- **Description**: Count the unique word pairs (bigrams) in the song's lyrics.
- **Calculation**: Generates counts of unique word pairs(bigrams) in the song's lyrics using the same custom N-gram parser.

10. trigrams

- **Description**: Count the unique wor triplets (trigrams) in the song's lyrics.
- **Calculation**: Generates counts of unique word triplets (trigrams) in the song's lyrics using the same custom N-gram parser

11. numberOfRepeatedWords

- **Description**: The count of words that are repeated in the song's lyrics.
- **Calculation**:Calculates the count of repeated words in the song's lyrics by subtracting the count of unique words from the total word count.

12. ratioOfTotalWordsToUnique

- **Description**: The ratio of unique words to total words in the song's lyrics.
- **Calculation**:Computes the ratio of unique words to total words in the song's lyrics.

13. percentageOfTotalWordsToUnique

- **Description**: The percentage of unique words out of the total words in the song's lyrics.
- **Calculation**: Calculates the percentage of unique words relative to the total words in the song's lyrics by multiplying the ratio of total words to unique words by 100.

14. LemmatizedWords

- **Description**: A list of lemmatized (base form) words from the song's lyrics.
- **Calculation**: Produces a list of lemmatized (base form) words from the song's lyrics using a text parsing module.

15. POSperWord

- **Description**: A list of parts of speech for each word in the song's lyrics.
- **Calculation**: Generates a list of parts of speech tags for each word in the song's lyrics using a POS tagging module.

16. sentimentScore

- **Description**: A numerical score representing the overall sentiment of the song's translated lyrics.
- **Calculation**: Provides a numerical score representing the overall sentiment of the song's translated lyrics using the Afinn sentiment analysis tool.

17. positiveWords

- **Description**: The count of positive words in the song's translated lyrics.
- **Calculation**: Counts the number of positive words in the song's translated lyrics using a sentiment analysis method that evaluates each word individually.

18. negativeWords

- **Description**: The count of negative words in the song's translated lyrics.
- **Calculation**: Counts the number of negative words in the song's translated lyrics using the same method as for positive words.

19. numberOfDiffLemmas

- **Description**: The number of different lemmas (base forms) in the song's lyrics.
- **Calculation**: Determines the number of different lemmatized forms in the song's lyrics.

20. numberOfDiffPOS

- **Description**: The number of different parts of speech in the song's lyrics.
- **Calculation**: Determines the number of different parts of speech tags in the song's lyrics

21. avgSetWordLength

- **Description**: The average length of unique words in the song's lyrics.
- **Calculation**: Calculates the average length of unique words in the song's lyrics.

22. avgAllWordLength

- **Description**: The average length of all words in the song's lyrics.
- **Calculation**: Calculates the average length of all words in the song's lyrics.

23. readabilityMeasure

- **Description**: A measure of the readability of the song's translated lyrics, calculated using readability formulas.
- **Calculation**:Computes a readability score for the song's translated lyrics using the Flesch and Fog readability formulas.

24. amountOfWordsRhymes

- **Description**: The number of rhyming words in the song's translated lyrics.
- **Calculation**: Counts the number of rhyming word pairs in the song's lyrics.(suffix 2)

25. ratioOfWordsToPOS

- **Description**: The ratio of different parts of speech to the total number of words in the song's lyrics.
- **Calculation**: Calculates the ratio of different parts of speech to the total number of words in the song's lyrics.

26. amountOfBiGrams

- **Description**: The number of unique bigrams (word pairs) in the song's lyrics.
- **Calculation**:Counts the number of unique bigrams in the song's lyrics.

27. amountOfTriGrams

- **Description**: The number of unique trigrams (word triplets) in the song's lyrics.
- **Calculation**: Counts the number of unique trigrams in the song's lyrics.

28. bigramsEntropy

- **Description**: The entropy (measure of randomness) of the distribution of bigrams in the song's lyrics.
- **Calculation**:Computes the entropy (a measure of randomness or diversity) for the distribution of bigrams in the song's lyrics.

29. trigramsEntropy

- **Description**: The entropy (measure of randomness) of the distribution of trigrams in the song's lyrics.
- **Calculation**:Computes the entropy for the distribution of trigrams in the song's lyrics.

30. avgSimilarityMeasure

- **Description**: The average semantic similarity between words in the song's lyrics.
- **Calculation**: Calculates the average semantic similarity between all pairs of words in the song's lyrics.

31. numberOfUniqueRankedWords

- **Description**: The number of unique words in the song's lyrics that are considered unique based on their frequency rank.

- **Calculation**: Counts the number of unique words in the song's lyrics that are rare based on their frequency rank.

32. avgUniquenessOfSong

- **Description**: The average uniqueness score of the words in the song's lyrics.
- **Calculation**:Calculates the average uniqueness score of the words in the song's lyrics based on their frequency.

33. repetitionWordsPercentage

- **Description**: The percentage of words that are repeated more than four times in the song's lyrics.
- **Calculation**:Computes the percentage of words that are repeated more than four times in the song's lyrics.

34. repetitionWordsUniqueness

- **Description**: The average uniqueness score of the words that are repeated more than four times in the song's lyrics.
- **Calculation**:Calculates the average uniqueness score of the words that are repeated more than four times in the song's lyrics.

35. semantic_similarity

- **Description**: Measures the cosine similarity between the embeddings of the original and translated lyrics.
- **Calculation Method**: Embeddings from BERT are calculated for both the original and translated lyrics. The cosine similarity between these embeddings is then computed.

36. average_word_frequency

- **Description**: Calculates the average frequency of all words in the song lyrics, based on a large word frequency list.
- **Calculation Method**: Each word in the lyrics is looked up in a frequency list to find its frequency of use in a given language. The average of these frequencies provides the average word frequency.

37. heBERT_sentiment

- **Description**: The sentiment score derived from the Hebrew BERT (heBERT) model.
- **Calculation Method**: The song lyrics are input into a heBERT model fine-tuned for sentiment analysis. The output score reflects the overall sentiment conveyed by the lyrics.

38. avg_word_similarity_hebrew

- **Description**: The average cosine similarity between all pairs of Hebrew word embeddings in the lyrics.
- **Calculation Method**: Hebrew words are converted into embeddings using a pre-trained model. Cosine similarity is calculated for every pair of embeddings, and the average is taken.

39. avg_word_similarity_english

- **Description**: The average cosine similarity between all pairs of English word embeddings in the translated lyrics.
- **Calculation Method**: Similar to Hebrew, but applies to the English translation of the lyrics.

40. Birth Year

- **Description**: The birth year of the artist or band.
- **Calculation Method**: Calculated by fetching the artist's Wikipedia page, extracting text surrounding the word "נולד," and then using a regular expression to locate and return the four-digit year found within that text segment.

41. word_similarity-large

- **Description**: A measure of the overall similarity between words in a song's lyrics, intended for longer texts.
- **Calculation Method**: Large text embeddings are generated for the entire lyrics, and similarity scores are calculated and averaged over the text span.

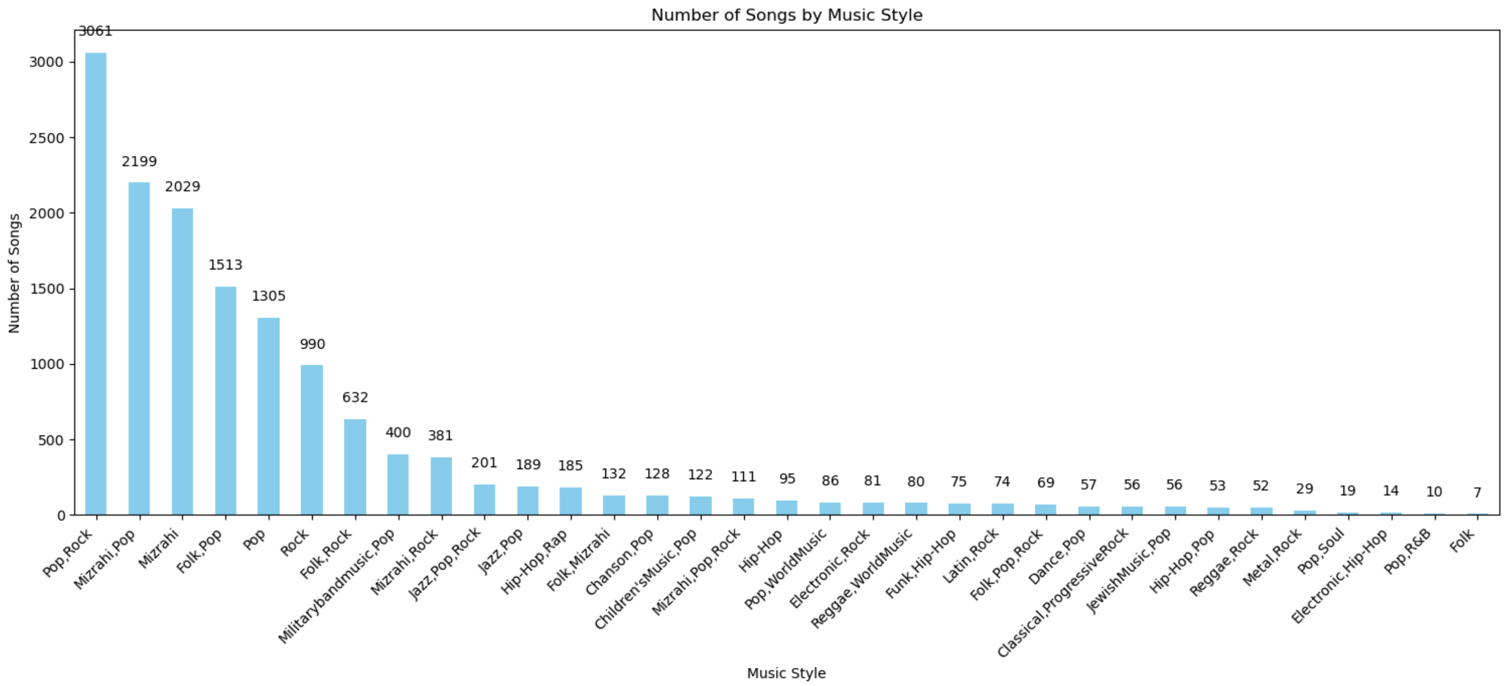
Table Data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Unname	name	artist	words	translate	Lemmati:	POSper\	songInEr	wordCou	uniqueW	releaseY	numberC	ratioOfTc	percenta
2	0	אחרי שאמ	אדם	אחרי שאמ	After I die	אחרי מת בא	ADP VEF	After I die	103	79	2019	24	0.767	76.699
3	1	אין יותר שז	אדם	אתה תראו	You'll see	ת הראה את	PRON VE	You'll see	117	85	2019	32	0.7265	72.65
4	2	אין מוצא	אדם	משוטט בר	Wanderin	משוטט ר	VERB NC	Wanderir	67	47	1987	20	0.7015	70.149
5	3	איש עוד לא	אדם	איש עוד לא	No one h	איש עוד לא	NOUN AI	No one h	134	48	1989	86	0.3582	35.821
6	4	אל תשכח	אדם	אנשים טוע	People m	יש טועה ח	NOUN VE	People m	135	61	1988	74	0.4519	45.185
7	5	אל תתני לי	אדם	בסוף תמונ	At the en	סוף תמונה	NOUN NC	At the en	93	73	2022	20	0.7849	78.495
8	6	ב' הקטנה	אדם	לא שמעתי	I haven't	לא שמע ע	ADV VEF	I haven't	130	46	1989	84	0.3538	35.385
9	7	באפריל או	אדם	הדרכים נצ	The road	דרך נצבע	NOUN VE	The road	162	91	2022	71	0.5617	56.173
10	8	בוקר טוב ל	אדם	בוקר טוב א	Good mo	בוקר טוב א	NOUN AI	Good mo	87	59	1999	28	0.6782	67.816
11	9	בחזרה לקו	אדם	בדרך אל ר	On the w	דרך אל גל	NOUN AI	On the w	66	57	2018	9	0.8636	86.364
12	10	בחיק הטבי	אדם	כאן בחיק ר	Here in n	כאן חיק טב	ADV NOI	Here in n	142	96	1995	46	0.6761	67.606
13	11	בית	אדם	פרקים שלנ	Whole ch	פרק שלם כ	NOUN AI	Whole ch	128	93	1998	35	0.7266	72.656
14	12	בלב אפריק	אדם	ריקוד אירוט	An erotic	ריקוד ארוטי	NOUN AI	An erotic	106	78	1989	28	0.7358	73.585
15	13	בלדה בין כ	אדם	הנוגה שלו	Venus se	נוגה שלח	NOUN VE	Venus se	100	81	1989	19	0.81	81
16	14	בסוף כולם	אדם	הוא גר שם	He lives t	הוא גר שם	PRON VE	He lives t	111	65	2019	46	0.5856	58.559
17	15	ברוש	אדם	אני ראיתי	I saw a cy	אני ראה ב	PRON VE	I saw a cy	74	46	1999	28	0.6216	62.162
18	16	ברחוב הנש	אדם	ברחוב הנש	On the S	רחוב נשמר	NOUN NC	On the S	102	67	2023	35	0.6569	65.686
19	17	הדברים הכ	אדם	מה עוד לא	What els	מה עוד לא	PRON AI	What els	148	104	2019	44	0.7027	70.27
20	18	היא רוקדת	אדם	בלילות הא	In the las	ליל אחרון	NOUN AI	In the las	114	74	1999	40	0.6491	64.912
21	19	הימים שאנ	אדם	הימים שאנ	The days	יום אני יחד	NOUN PF	The days	92	73	1999	19	0.7935	79.348
22	20	הכרזת מל	אדם	נמאס מאוו	I'm tired c	נמאס את	VERB PF	I'm tired c	87	71	1987	16	0.8161	81.609
23	21	הנשיקה הו	אדם	אחותי הגד	My older	אחות גדול	NOUN AI	My older	166	88	2021	78	0.5301	53.012
24	22	הרוח	אדם	האם כדאי	Is it worth	אם כדאי ה	ADV AUX	Is it worth	81	41	1988	40	0.5062	50.617
25	23	ואם את עו	אדם	אם את עו	If you lea	אם את עזו	SCONJ F	If you lea	98	63	2018	35	0.6429	64.286
26	24	ואנחנו	אדם	תעלה השנ	The sun	עלה שמש	VERB NC	The sun	89	61	2023	28	0.6854	68.539
27	25	ונתנה תוקף	אדם	נתנה תוקף	Give effe	תוקף קדוו	VERB NC	Give effe	199	155	2022	44	0.7789	77.889
28	26	זה הזמן לו	אדם	אחרי שסיי	After I'm ANK]	אחרי	ADP VEF	After I'm	97	56	2022	41	0.5773	57.732
29	27	זה לא מסו	אדם	יש ימים שנ	There are	יש יום בא	VERB NC	There are	111	58	2021	53	0.5225	52.252
30	28	זריחה על	אדם	זריחה על	A sunrise	זריחה על	NOUN AI	A sunrise	89	67	1987	22	0.7528	75.281
31	29	חדר משלי	אדם	לרגל השקו	On the o	רגל השקה	ADP NOI	On the o	171	95	2021	76	0.5556	55.556
32	30	חורף	אדם	גשר צר שו	A narrow	גשר צר מ	NOUN AI	A narrow	85	55	2021	30	0.6471	64.706
33	31	חייל של גו	אדם	היום אני פו	Today I o	יום אני פותו	ADV PRC	Today I o	218	82	1988	136	0.3761	37.615
34	32	יחד נעמוד	אדם	אנחנו העם	We are th	אני עם הל	PRON NC	We are th	96	74	1991	22	0.7708	77.083
35	33	ימים ברקיע	אדם	אל תהפכי	Don't turn	אל התהפך	ADP VEF	Don't turn	88	66	1999	22	0.75	75
36	34	יש אהבה	אדם	זאת אולי ר	This may	זה אולי אה	PRON AI	This may	101	40	2020	61	0.396	39.604
37	35	כי לא אמר	אדם	לפעמים אנ	Sometim	לפעמים אנ	ADV PRC	Sometim	95	69	1999	26	0.7263	72.632
38	36	כמו גיבור	אדם	כמו פנתר	Like a ca	כמו פנתר	ADP NOI	Like a ca	172	116	1989	56	0.6744	67.442
39	37	לא לוקח ב	אדם	ברגע אחרו	At the las	רגע אחרון	NOUN AI	At the las	121	60	1987	61	0.4959	49.587
40	38	לא נגמר ה	אדם	לא נגמר ה	The sea	לא נגמר ים	ADV VEF	The sea	121	69	2021	52	0.5702	57.025
41	39	לא נוגע בך	אדם	אם לא אס	If I don't t	אם לא סיפ	SCONJ A	If I don't t	65	29	2022	36	0.4462	44.615
42	40	לא ניתן לך	אדם	אם אתה ב	If you're i	אם הוא חד	SCONJ F	If you're i	100	55	2021	45	0.55	55
43	41	לא נרדם	אדם	לא נרדם ל	I don't fal	לא נרדם ל	ADV VEF	I don't fal	105	67	1987	38	0.6381	63.81
44	42	לגמרי לא	אדם	איש קטן מ	little man	איש קטן מ	NOUN AI	little man	98	49	2021	49	0.5	50

	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	DiffLemm	DiffPOS	numberC	numberC	bigramsE	trigramsE	sentimer	average	WordsRh	RatioOfF	readabilit	positive	negative	avgSimila	Music St
2	73	10	93	99	6.4209	6.6186	-15	4.4937	189	0.0971	12.64	2	7	20.67	Pop
3	77	10	105	113	6.6024	6.8107	9	4.0824	66	0.0855	67.945	7	4	32.176	Pop
4	47	7	54	59	5.5292	5.8378	-14	4.0851	17	0.1045	45.84	2	13	24.114	Pop
5	42	8	70	85	5.6497	6.1069	-9	4.1042	116	0.0597	9.715	9	20	36.643	Pop
6	52	9	73	79	5.559	5.9198	-14	4.0164	124	0.0667	68.915	13	38	22.596	Pop
7	66	8	87	88	6.3985	6.4336	2	4.0274	51	0.086	10.64	3	3	31.524	Pop
8	44	9	58	62	5.6751	5.7754	-2	3.9348	271	0.0692	4.535	0	2	23.61	Pop
9	83	10	110	118	6.5314	6.6824	9	4.1538	93	0.0617	27.475	5	3	27.008	Pop
10	56	9	74	79	6.1064	6.2505	7	3.9831	30	0.1034	7.47	5	3	29.488	Pop
11	51	9	63	64	5.9492	6	-3	3.9474	28	0.1364	3.295	1	2	21.361	Pop
12	91	12	111	116	6.6302	6.7217	16	4.0521	125	0.0845	82.535	12	2	25.192	Pop
13	79	9	120	124	6.8441	6.9395	-4	3.7742	93	0.0703	12.96	7	8	25.416	Pop
14	75	7	88	94	6.2571	6.4504	-11	4.6282	111	0.066	75.035	1	6	18.488	Pop
15	71	11	92	94	6.4651	6.5254	6	4	93	0.11	55.74	7	4	20.184	Pop
16	65	9	77	79	6.0579	6.1376	16	4.1385	411	0.0811	10.43	9	3	33.281	Pop
17	44	9	59	63	5.7411	5.9094	-3	3.5	40	0.1216	16.05	4	6	27.683	Pop
18	60	8	77	79	6.1434	6.2039	5	4.7164	477	0.0784	59.7	8	4	18.416	Pop
19	96	10	118	119	6.8051	6.82	-5	4.4904	408	0.0676	11.14	4	8	21.432	Pop
20	70	9	95	103	6.3643	6.5974	-16	4.3378	95	0.0789	58.255	6	13	24.796	Pop
21	64	9	85	87	6.3676	6.4252	16	4.1781	102	0.0978	49.265	7	0	34.469	Pop
22	62	11	84	84	6.3798	6.3859	-2	4.0563	32	0.1264	53.34	8	12	26.354	Pop
23	81	11	118	128	6.5748	6.7839	15	4.1477	186	0.0663	10.48	12	7	20.461	Pop
24	38	10	48	53	5.3092	5.5252	15	3.7073	34	0.1235	46.44	11	7	25.154	Pop
25	59	11	80	86	6.1682	6.3401	-13	3.9524	50	0.1122	9.56	1	11	31.368	Pop
26	59	8	73	77	5.936	6.1164	8	4.1148	62	0.0899	20.04	5	1	35.394	Pop
27	131	9	194	196	7.589	7.6119	-26	4.0645	134	0.0452	10.02	11	17	28.518	Pop
28	51	9	72	76	6.0118	6.146	-7	4.375	152	0.0928	57.425	3	7	30.46	Pop
29	54	8	83	91	6.1756	6.3874	-2	3.7931	77	0.0721	13.64	7	10	36.464	Pop
30	64	6	77	79	6.1382	6.2417	11	4.3134	52	0.0674	48.94	11	4	26.358	Pop
31	86	12	122	126	6.8285	6.892	-9	4.0632	312	0.0702	38.8	5	12	27.024	Pop
32	53	8	67	68	5.9876	6.0136	-6	3.6909	48	0.0941	49.45	4	9	17.948	Pop
33	73	8	101	107	6.1642	6.3496	-19	4.2805	416	0.0367	14.595	2	14	17.545	Pop
34	65	9	90	92	6.4266	6.512	13	4.1081	153	0.0938	60.825	11	3	36.3	Pop
35	56	9	79	81	6.259	6.31	8	3.9545	42	0.1023	46.41	4	1	24.465	Pop
36	37	9	53	61	5.5212	5.7809	20	3.825	123	0.0891	10.235	11	5	38.941	Pop
37	64	9	84	87	6.2993	6.3886	-3	4.0145	85	0.0947	4.62	2	4	28.087	Pop
38	111	12	140	144	7.042	7.0946	23	4.0259	216	0.0698	83.515	22	10	25.348	Pop
39	57	8	72	75	5.8724	5.9768	-19	4.3667	302	0.0661	5.47	5	14	19.392	Pop
40	63	7	84	91	6.2566	6.3989	4	4.1014	80	0.0579	19.745	5	6	22.652	Pop
41	27	7	42	51	5.2105	5.5364	3	3.6897	13	0.1077	1.96	1	0	21.034	Pop
42	53	9	74	84	5.9573	6.2905	13	3.9455	14	0.09	6	9	4	29.348	Pop
43	65	10	82	89	6.2354	6.4147	5	3.8209	83	0.0952	4.945	4	3	16.95	Pop
44	46	8	70	78	5.8588	6.133	2	4.3469	66	0.0816	47.935	9	6	22.496	Pop

	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO
1	NumberC	AvgUniqu	percenta	theUniqu	MappedS	semantic	average_	heBERT_	avg_wor	avg_wor	Birth_Ye	word_similarity-la
2	13	1E+06	1.9417	170079	0	0.6638	0.0009	-1	0.322	0.3672	1960	0.8595
3	7	418108	2.5641	4573.2	0	0.6924	0.0014	-1	0.4392	0.3152	1960	0.8697
4	6	290931	2.9851	5316.9	0	0.7588	0.0009	-1	0.4172	0.3677	1960	0.8636
5	13	718666	4.4776	1257.5	0	0.5561	0.0023	-1	0.4389	0.3636	1960	0.8804
6	2	94938	1.4815	20627	0	0.5794	0.0013	-1	0.4404	0.3448	1960	0.8765
7	3	170971	1.0753	436.68	0	0.6901	0.0023	-1	0.4539	0.4136	1960	0.8756
8	2	219108	5.3846	2864.8	0	0.6416	0.0033	-1	0.4476	0.293	1960	0.8678
9	17	513853	3.0864	27200	0	0.6003	0.0027	-1	0.3229	0.3493	1960	0.8634
10	4	116944	0	0	0	0.6873	0.0023	-1	0.4385	0.4082	1960	0.869
11	1	88376	0	0	0	0.6618	0.0014	-1	0.3496	0.3586	1960	0.8669
12	9	273253	1.4085	109.43	0	0.6551	0.0017	-1	0.2946	0.3138	1960	0.8632
13	8	1E+06	1.5625	437.37	0	0.6796	0.0029	-1	0.4666	0.3064	1960	0.8671
14	12	624318	1.8868	17451	0	0.64	0.0008	-1	0.3715	0.3328	1960	0.8555
15	15	1E+06	0	0	0	0.6689	0.0017	-1	0.4498	0.2878	1960	0.8599
16	3	92588	5.4054	2703.8	0	0.567	0.0015	-1	0.4234	0.3912	1960	0.8741
17	6	820119	2.7027	101241	0	0.6335	0.002	-1	0.405	0.4244	1960	0.8591
18	12	293185	0.9804	70.922	0	0.4355	0.0015	-1	0.326	0.3335	1960	0.8571
19	16	893112	0.6757	49.02	0	0.6495	0.002	-1	0.3987	0.3076	1960	0.8657
20	8	2E+06	3.5088	60389	0	0.5905	0.0029	-1	0.3213	0.3371	1960	0.8701
21	3	161741	1.087	165.84	0	0.7148	0.0025	-1	0.4451	0.4056	1960	0.8759
22	4	156850	1.1494	61.728	0	0.7348	0.0029	-1	0.4699	0.3274	1960	0.8714
23	6	648805	3.012	382.03	0	0.6027	0.0017	-1	0.2963	0.3152	1960	0.8682
24	4	516394	3.7037	2714.3	0	0.682	0.0011	-1	0.4578	0.3644	1960	0.8749
25	7	1E+06	3.0612	174.66	0	0.6636	0.0036	-1	0.4276	0.3499	1960	0.8702
26	6	670098	3.3708	2428.6	0	0.7053	0.0014	-1	0.4552	0.3978	1960	0.8661
27	37	2E+06	1.005	476.9	0	0.5491	0.0012	-1	0.3423	0.3702	1960	0.857
28	5	194712	4.1237	19920	0	0.71	0.0025	-1	0.3941	0.3104	1960	0.8754
29	0	18545	3.6036	873.09	0	0.6376	0.0036	-1	0.4425	0.3524	1960	0.8808
30	3	172647	2.2472	7513.3	0	0.7403	0.0014	-1	0.4044	0.3573	1960	0.8623
31	5	337466	1.1696	83.22	0	0.6348	0.0023	-1	0.3052	0.2822	1960	0.8697
32	5	307216	2.3529	23573	0	0.6653	0.0017	-1	0.3633	0.3677	1960	0.8634
33	17	344226	5.0459	18743	0	0.527	0.0028	-1	0.2918	0.3195	1960	0.8682
34	4	484466	3.125	70251	0	0.6666	0.0013	-0.606	0.4691	0.4095	1960	0.8701
35	5	1E+06	1.1364	49.02	0	0.7108	0.0024	-1	0.4635	0.3497	1960	0.8653
36	1	79059	2.9703	890.05	0	0.6698	0.0024	-1	0.4388	0.3718	1960	0.8738
37	4	544959	2.1053	275.96	0	0.6944	0.0025	-1	0.4321	0.3805	1960	0.8731
38	14	533306	1.7442	144986	0	0.6636	0.0013	-1	0.4024	0.2912	1960	0.8649
39	10	717658	2.4793	9280.6	0	0.6312	0.0033	-1	0.3386	0.3176	1960	0.8672
40	18	2E+06	3.3058	167.97	0	0.4721	0.0028	-1	0.3939	0.3372	1960	0.8581
41	3	374081	3.0769	11473	0	0.6938	0.0037	-1	0.3125	0.3865	1960	0.8781
42	4	811288	4	567.69	0	0.6803	0.0025	-1	0.4242	0.3263	1960	0.8798
43	1	148445	1.9048	67507	0	0.697	0.0031	-1	0.391	0.3158	1960	0.8746
44	9	334355	6.1224	328481	0	0.6131	0.003	-1	0.3843	0.3178	1960	0.8746

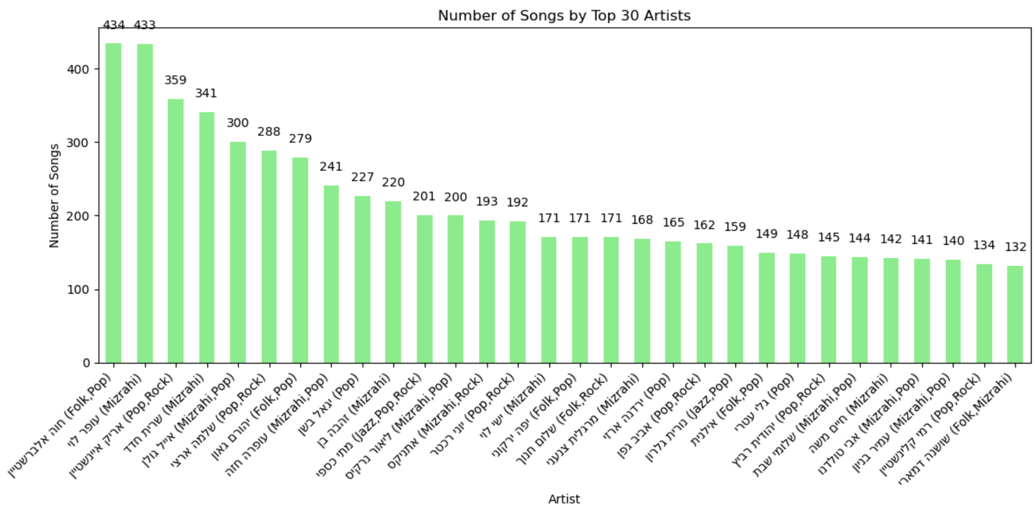
Over look on the Data:



The five primary music styles represented in the data are:

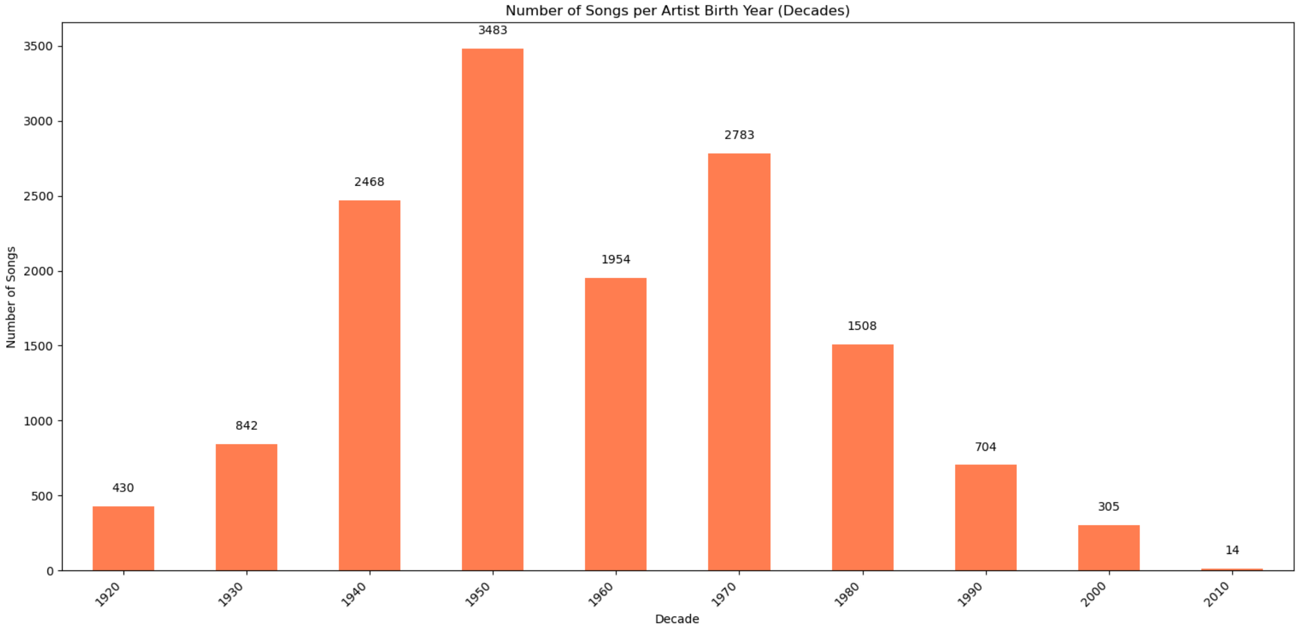
- Pop & Rock
- Mizrahi & Pop
- Mizrahi
- Folk & Pop
- Pop

Most of the data is related to Pop music, either by itself or in combination with other styles.

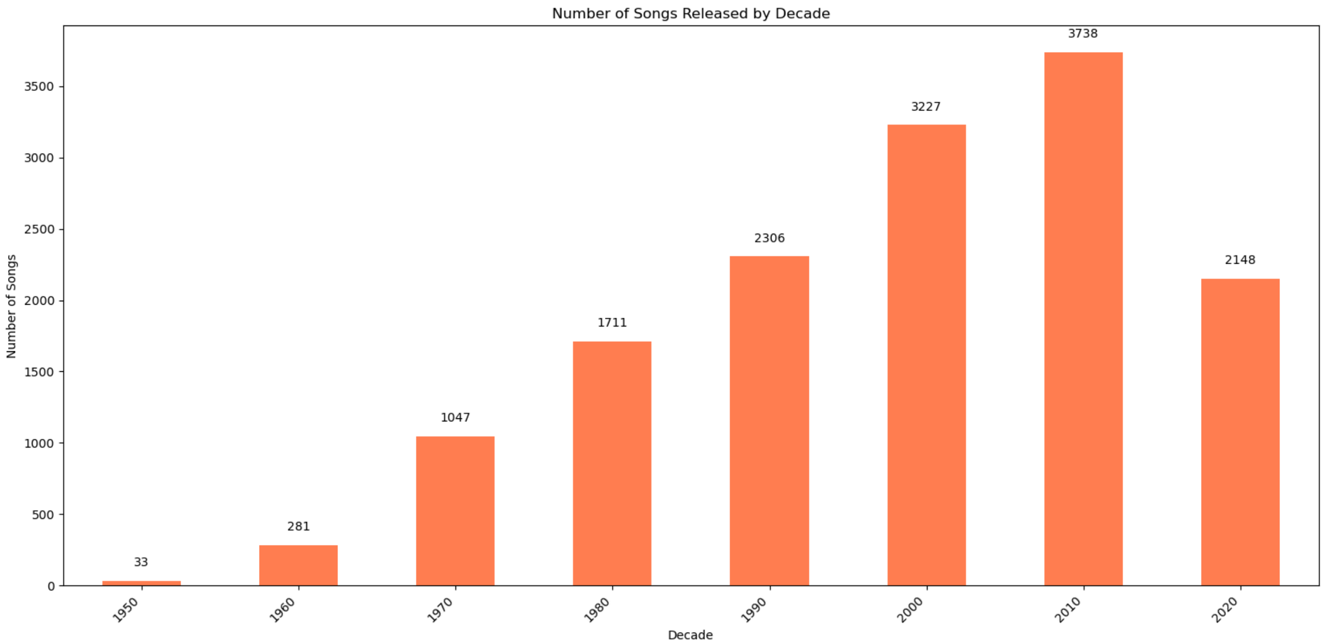


Top 30 artists by song count show a variety of artists and music styles. The highest count of songs in the dataset is 434 by "חוה אלברשטיין" and 433 by "עופר לוי", indicating these artists' significant contributions. This diversity suggests a rich dataset with substantial representation from prominent figures in the music scene.

Most of the artists in the dataset were born between 1940 and 1990, with a significant concentration in the 1950s and 1960s.



The majority of songs in the dataset were released between 1980 and 2023.



ML Prediction

We aimed to determine if different music styles can be classified solely based on numerical features extracted from song lyrics and other metadata. The process involved the following steps:

1. **Train-Test Split**: The dataset was divided into two sets, with 80% used for training the model and 20% for testing it. This split was stratified to ensure equal representation of each genre in both sets.
2. **Train a Classifier**: A RandomForestClassifier, a machine learning model, was trained on the training set. This model learned to classify songs into different genres based on various extracted features.
3. **Predict and Evaluate**: The trained model was used to predict the genres of the songs in the test set. The accuracy of these predictions was calculated for each genre, providing a measure of the model's performance.
4. **Feature Importance**: The importance of each feature in making predictions was calculated and ranked. This analysis helped identify which features were most influential in distinguishing between genres, providing insights into the characteristics that define each genre.

Hypothesis Statement: We hypothesize that the **"Folk,Pop"** genre can be distinguished from the **"Mizrahi,Pop"** genre based on distinct lexical and structural features in their lyrics. Specifically, we expect "Folk,Pop" songs to exhibit greater lexical diversity and complexity, while "Mizrahi,Pop" songs will demonstrate higher word similarity and sentiment-oriented readability.

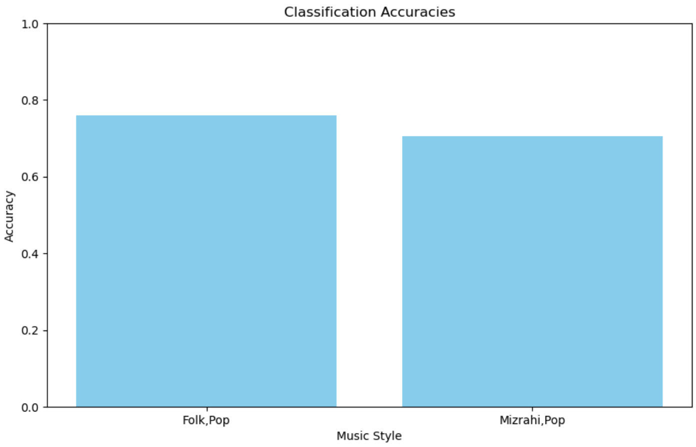
- **Lexical Diversity and Complexity** in "Folk,Pop": We expect "Folk,Pop" songs to have higher values in features like unique words, bi-grams, and tri-grams, indicating more complex and varied lyrical structures.

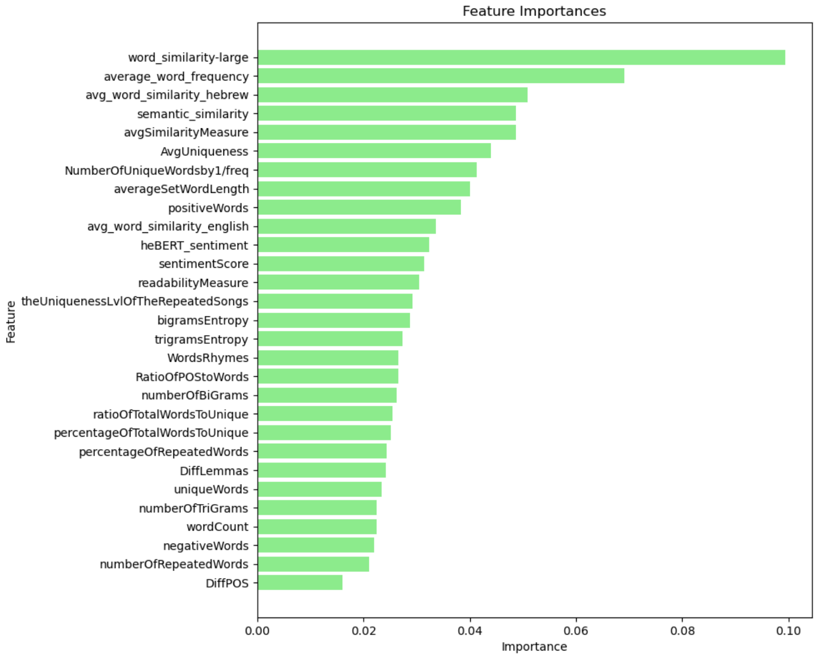
- **Higher Word Similarity** in "Mizrahi,Pop": The `word_similarity-large` feature should be higher in "Mizrahi,Pop" songs, suggesting that the words used are more similar to each other within songs.

- **Common Words** in "Mizrahi,Pop": The average_word_frequency feature is anticipated to be higher in "Mizrahi,Pop" songs, indicating the use of more common, less unique words.

- **Sentiment and Readability** in "Mizrahi,Pop":*We hypothesize that "Mizrahi,Pop" songs will show higher sentiment scores and readability measures, reflecting a focus on emotional expression and simpler, more accessible language.

The results of the model's predictions confirm the hypothesis, achieving good accuracy for both genres around 0.74. The important distinguishing features were word_similarity-large, average_word_frequency, and avg_word_similarity_hebrew. These findings suggest that "Mizrahi,Pop" songs use more similar and common words, while "Folk,Pop" songs exhibit more lexical diversity and complexity. Additionally, "Mizrahi,Pop" songs are characterized by higher sentiment scores and readability, indicating a focus on sentiment and simpler readability.



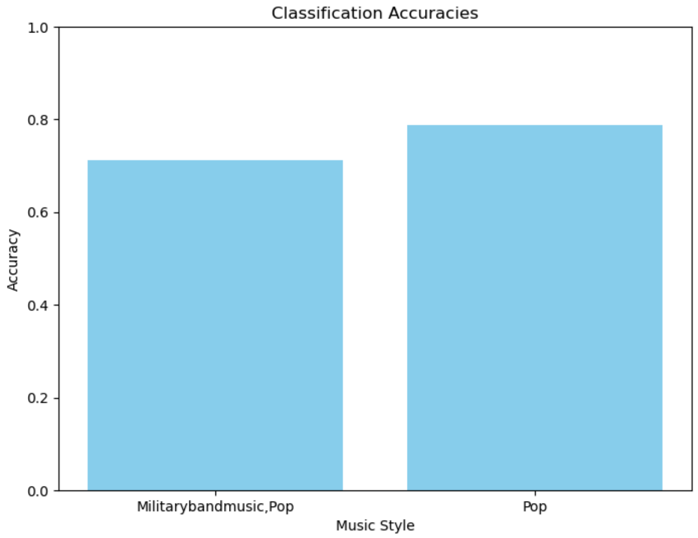


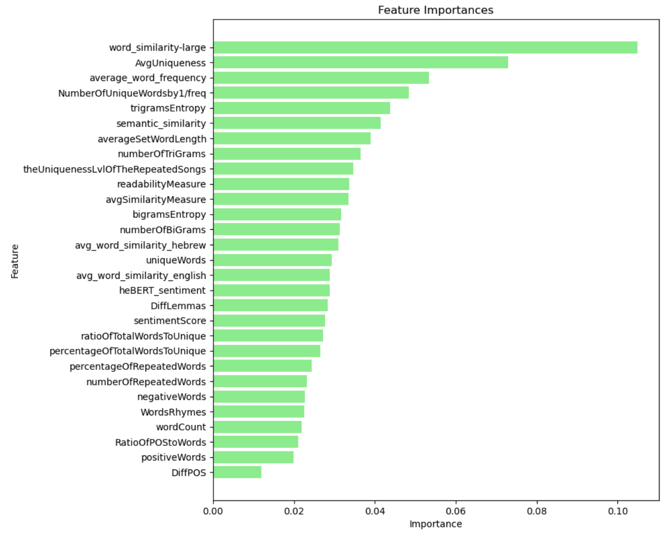
Hypothesis Statement: We hypothesize that the "Militarybandmusic,Pop" genre can be distinguished from the "Pop" genre based on distinct lexical and structural features in their lyrics. Specifically, we expect "Militarybandmusic,Pop" songs to exhibit greater lexical diversity and complexity, while "Pop" songs will demonstrate higher word similarity, sentiment-oriented readability, and the use of more common words.

- **Lexical Diversity and Complexity in "Militarybandmusic,Pop"**: We expect "Militarybandmusic,Pop" songs to have higher values in features like unique words, bi-grams, and tri-grams, indicating more complex and varied lyrical structures.
- **Higher Word Similarity in "Pop":** The `word_similarity-large` feature should be higher in "Pop" songs, suggesting that the words used are more similar to each other within songs.
- **Common Words in "Pop":** The average_word_frequency feature is anticipated to be higher in "Pop" songs, indicating the use of more common, less unique words.
-**Sentiment and Readability in "Pop":** We hypothesize that "Pop" songs will show higher sentiment scores and readability measures, reflecting a focus on emotional expression and simpler, more accessible language.

Results

The classification results **support** the hypothesis, showing an accuracy of 71% for predicting "Militarybandmusic,Pop" and 79% for predicting "Pop". The important distinguishing features were `word_similarity-large`, `AvgUniqueness`, `average_word_frequency`, and `NumberOfUniqueWordsby1/freq`. These findings suggest that "Militarybandmusic,Pop" songs use more unique and higher-frequency words, and exhibit more complex lyrical structures. In contrast, "Pop" songs are characterized by higher word similarity, sentiment scores, readability measures, and the use of more common words, indicating a focus on sentiment and simpler readability.





Hypothesis Statement: We hypothesized that **"Folk,Pop"** and **"Folk,Rock"** genres would be distinguishable based on lexical and structural features, expecting "Folk,Pop" songs to have distinctively higher positivity and more common word usage compared to "Folk,Rock" songs.

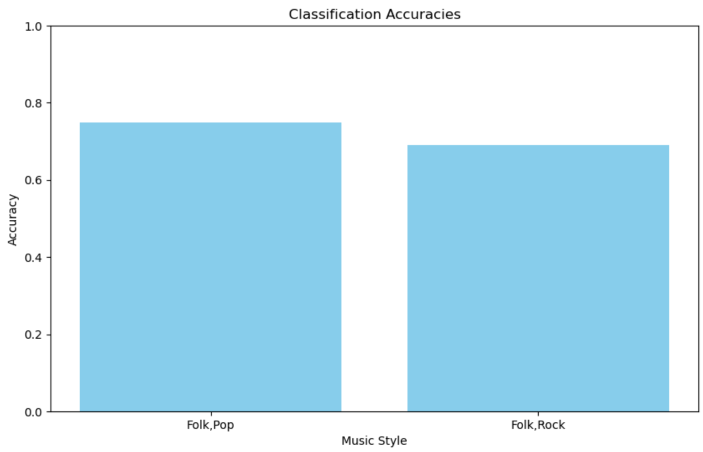
Expected Results

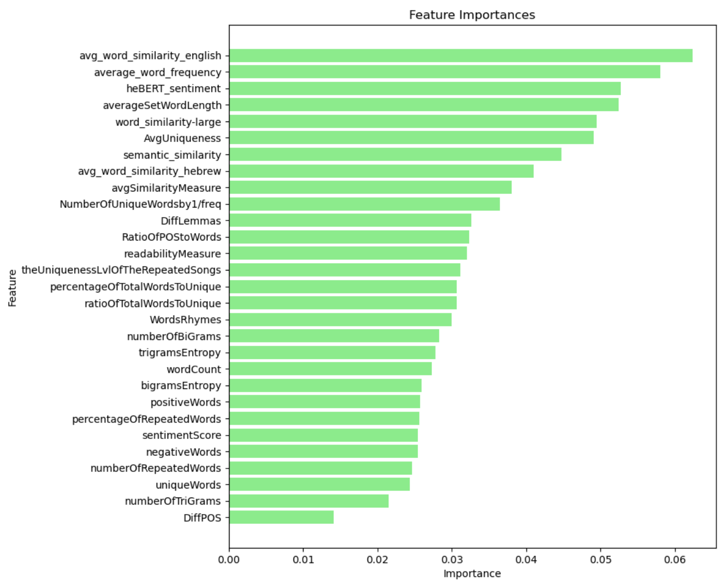
1. **higher Word Similarity in "Folk,Pop":** Smaller variety of words used.
2. **Common Word Usage in "Folk,Pop":**Use of more common words.
3. **Higher Positivity in "Folk,Pop":** More positive sentiment in lyrics.

Actual Results

1. **Higher Word Similarity in "Folk,Rock":** Words tend to be more similar to each other.
2. **Higher Common Word Usage in "Folk,Rock":**Use of more common words.
3. **Higher Positivity in "Folk,Pop":** More positive sentiment, aligning with our hypothesis.

Conclusion: The results **did not support** our hypothesis. While we expected distinctions based on our selected features, the actual classification showed that most of these features(except the positivity) did not consistently align with our expectations.





For artist classification, the same methodology was applied, allowing us to distinguish between different artists based on their features.
In the classification task, we aimed to distinguish between **different artists** based on various song features. The overall classification accuracy for each artist and the importance of various features in making these predictions were analyzed. Here are the key findings:

Hypothesis Statement: Inspired by the results of the genre classification hypothesis, we hypothesize that **"יהורם גאון"** ("Folk, Pop") and **"אייל גולן"** ("Mizrahi, Pop") can be accurately classified based on features such as lexical diversity, word similarity, and word frequency, with "יהורם גאון" exhibiting more lexical diversity and less word similarity compared to "אייל גולן".

Expected Results

1. **Higher Lexical Diversity** for "יהורם גאון":*More unique words and varied vocabulary.

2. **Lower Word Similarity** for "יהורם גאון":*Less similarity between words within songs.

3. **Higher Common Word Usag**e for "אייל גולן": More frequent use of common words.

4. **Higher Sentiment Scores and Readability** for "אייל גולן":*More positive sentiment and easier readability.

Actual Results

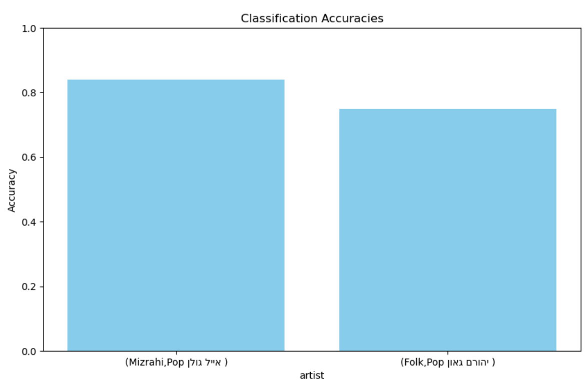
1. **Higher Lexical Diversit**y for "יהורם גאון":Confirmed by the feature NumberOfUniqueWordsby1/freq.

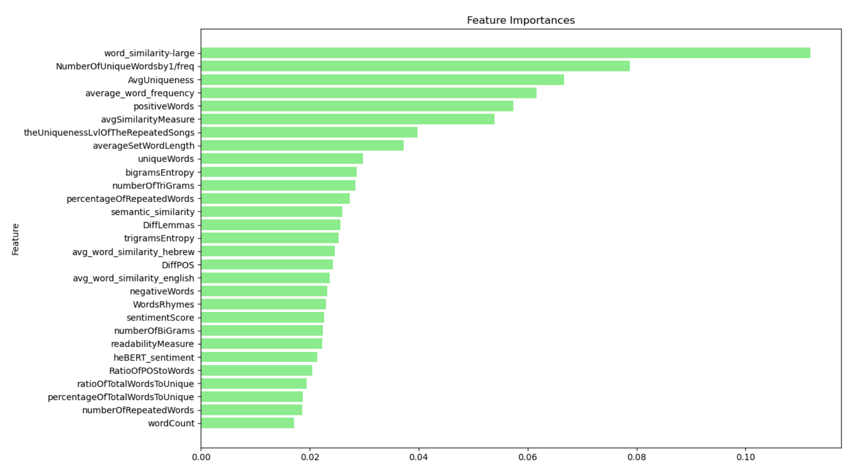
2. **Lower Word Similarity** for "יהורם גאון": Confirmed by the feature word_similarity-large.

3. **Higher Common Word Usag**e for "אייל גולן": Confirmed by average_word_frequency.

4. **Higher Sentiment Scores and Readability** for "אייל גולן": Confirmed by the model's predictions.

Conclusion: The model's predictions **confirm the hypothesis**, achieving good accuracy for both artists with around 80% successful predictions. The important distinguishing features were word_similarity-large, NumberOfUniqueWordsby1/freq, and average_word_frequency. These findings suggest that "יהורם גאון" has more lexical diversity and less word similarity within his songs, while "אייל גולן" uses more common words and exhibits higher sentiment scores and readability. Thus, the hypothesis that these features can effectively classify songs of these specific artists is supported by the results.





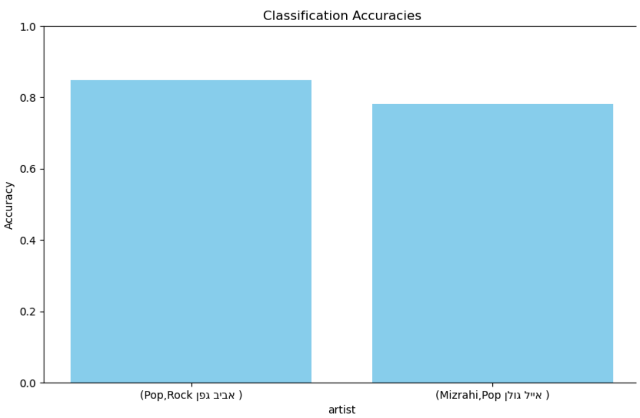
Hypothesis:"אייל גולן" (Mizrahi&Pop) songs are characterized by more positive sentiment, less predictable lyrical structures, and greater complexity compared to"אביב גפן"(Pop&Rock) songs. In contrast, Aviv Geffen's lyrics show higher lexical diversity.

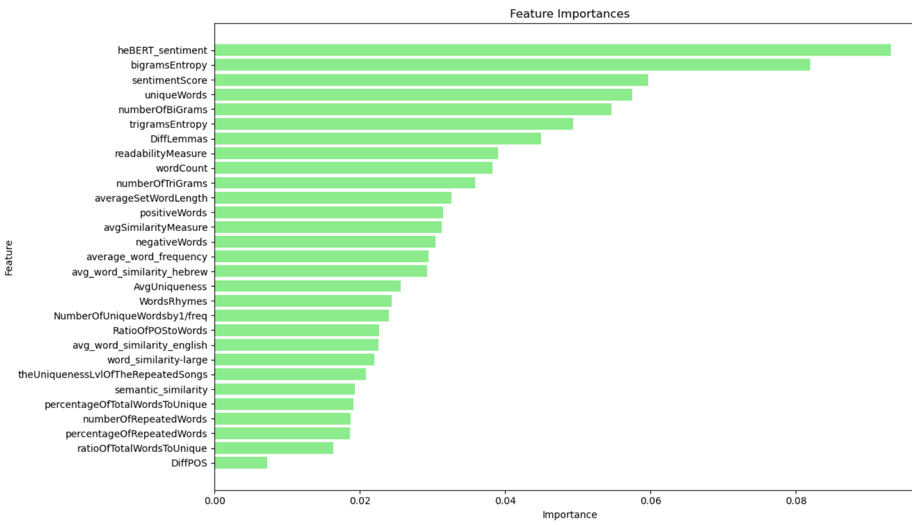
Results: The model's predictions **confirm the hypothesis** The classification results between **Aviv Geffen** and **Eyal Golan** show high accuracy, with Geffen at 85% and Golan at 78%.

- **Sentiment:** Eyal Golan's lyrics tend to be more positive.

- **Predictability:** Golan's lyrics are less predictable (higher bigrams entropy), indicating greater complexity.

- **Lexical Diversity:** Aviv Geffen's lyrics have a higher ratio of total words to unique words, reflecting greater lexical diversity.

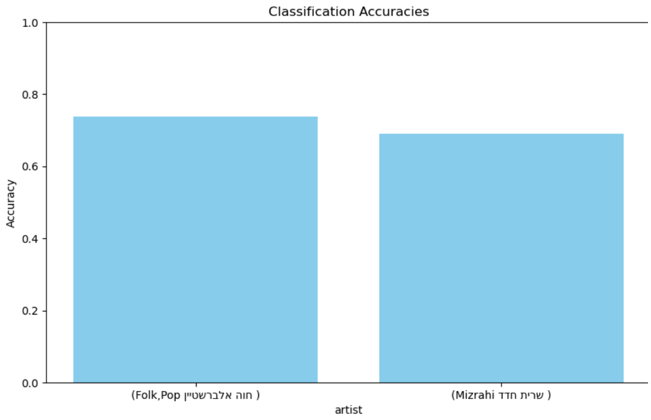


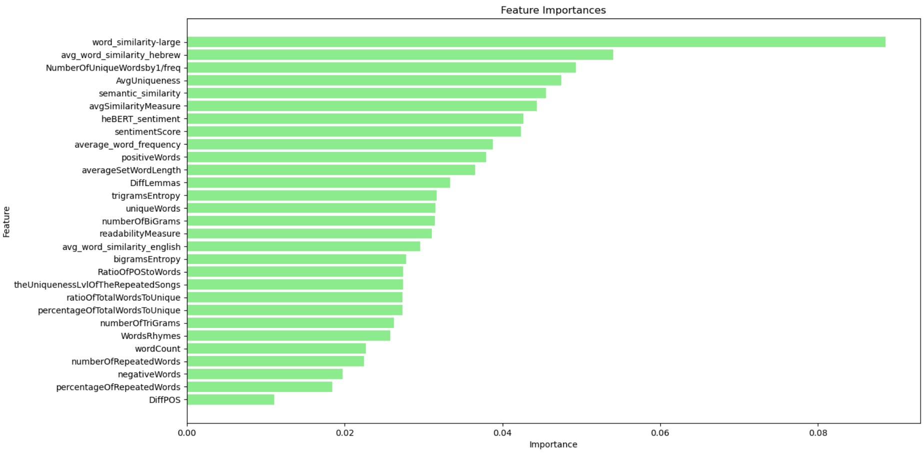


Hypothesis: **"חוה אלברשטיין"**(Folk,Pop) lyrics are characterized by greater linguistic diversity and unique word usage, while **"שרית חדד"**(Mizrahi) lyrics display higher word similarity and thematic cohesion.

Results: The classification between artists **"חוה אלברשטיין"** and **"שרית חדד" confirm the hypothesis** shows accuracies of 74% and 69%, respectively. The most important features for distinguishing their songs are `word_similarity-large`, `avg_word_similarity_hebrew`, `NumberOfUniqueWordsby1/freq`, and `AvgUniqueness`.

- **Linguistic Diversity**: Hava Alberstein's lyrics have higher unique words and greater linguistic diversity.
- **Word Similarity**: Sarit Hadad's lyrics tend to use words that are semantically similar or fall within the same lexical fields indicating more cohesive and thematically focused content.

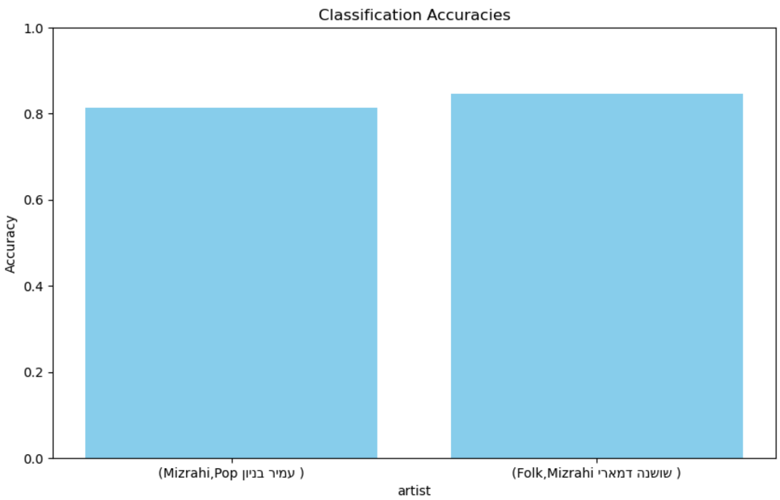


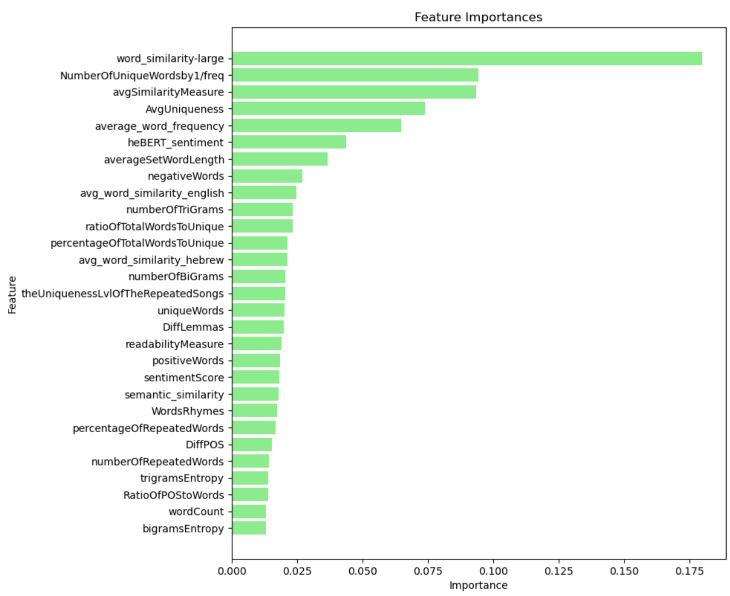


Hypothesis: "**שושנה דמארי**"(Folk&Mizrahi) lyrics exhibit greater lexical richness and diversity, while"**עמיר בניון**"(Mizrahi&Pop) lyrics show higher word similarity.

Results: The classification resulted in an accuracy of 81% for "עמיר בניון" and 85% for "שושנה דמארי". **confirm the hypothesis:**

- **Lexical Richness**: Shoshana Damari's lyrics feature higher word count, unique words, and bi-grams, reflecting greater lexical diversity.
- **Word Similarity**: Amir Benayoun's lyrics have higher word similarity, indicating more consistent use of similar words within his songs.





Hypothesis Statement: We hypothesize that predicting a singer from the same genre will result in a low percentage of success due to the inherent similarities in lyrical and musical styles within the same genre.

1.we compare between the “Mizrahi” genre **“חיים משה”& ”ישי לוי”.**

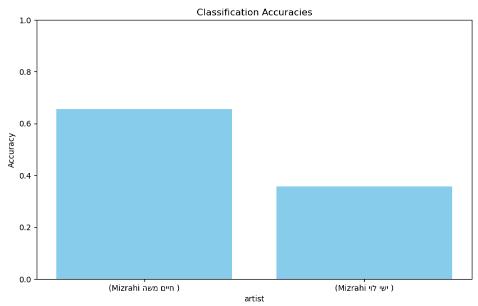
- **חיים משה Accuracy: 0.66**
- **ישי לוי Accuracy: 0.36**

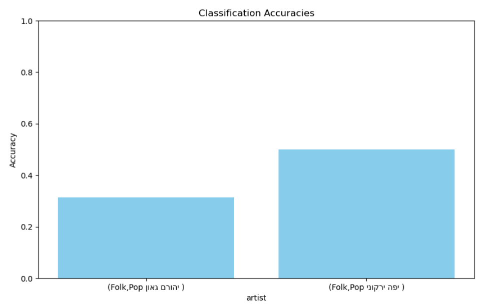
2. we compare between “Folk,Pop” genre -”יהורם גאון”&”יפה ירקוני”

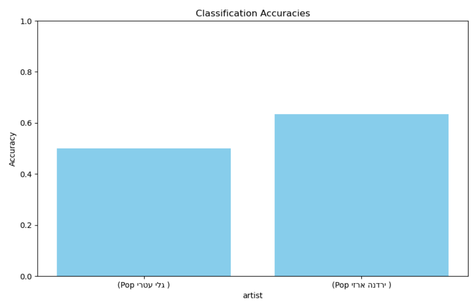
- **יהורם גאון Accuracy: 0.31**
- **יפה ירקוני Accuracy: 0.50**

3. We compared between “Pop” genre- “ירדנה ארזי”&”גלי עטרי”

- **גלי עטרי Accuracy: 0.57**
- **ירדנה ארזי Accuracy: 0.53**







Conclusion: The hypothesis **is supported** by the results, as the lower accuracy in predicting highlights the difficulty in differentiating artists within the same genre. The similarities in lyrical content, musical style, and thematic elements likely contribute to this challenge.

Creativity Measure

Feature Selection and Inversion:

- A set of features related to lyrical creativity:Some features are inverted to align with the creativity score, meaning higher values in these features represent higher creativity.

```
adjusted_creativity_features = [
    'uniqueWords', 'ratioOfTotalWordsToUnique', 'percentageOfTotalWordsToUnique',
    'DiffLemmas', 'DiffPOS', 'bigramsEntropy', 'trigramsEntropy',
    'averageSetWordLength', 'WordsRhymes', 'RatioOfPOStoWords','NumberOfUniqueWordsby1/freq',
    'inv_avgSimilarityMeasure', 'inv_average_word_frequency','inv_avg_word_similarity_hebrew','inv_avg_word_similarity_english'
]
```

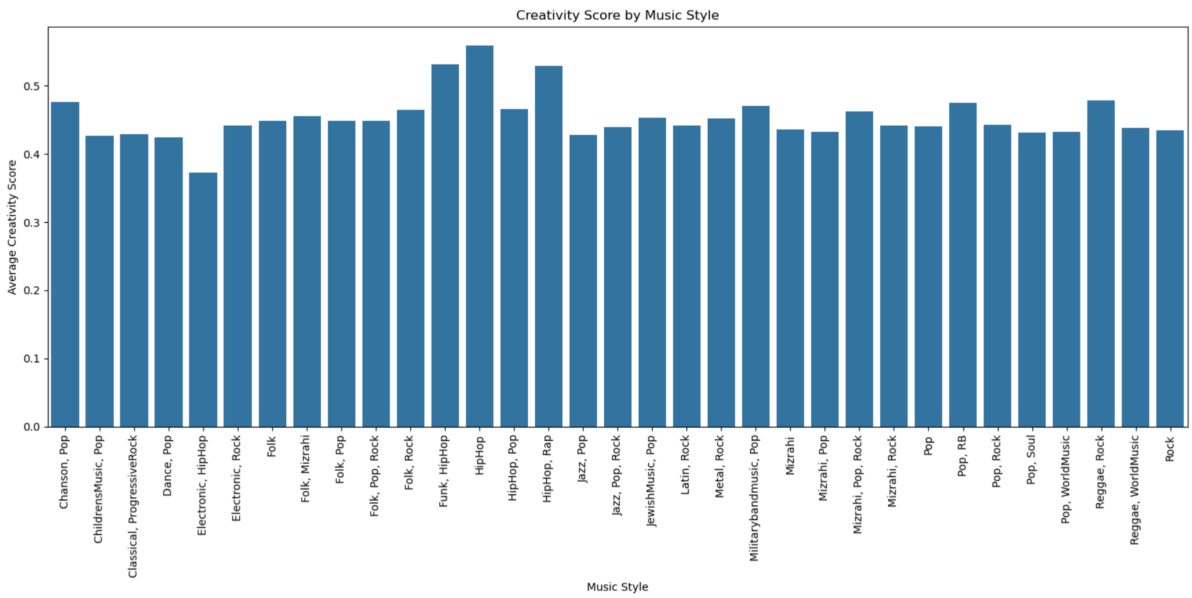
Normalization:

- The selected features are normalized to a range of 0 to 1 using `MinMaxScaler` to ensure comparability across different features.

Creativity Score Calculation:

- The normalized features are combined to create a single creativity score for each song.
- The creativity score is aggregated over the years and by music style to analyze trends.

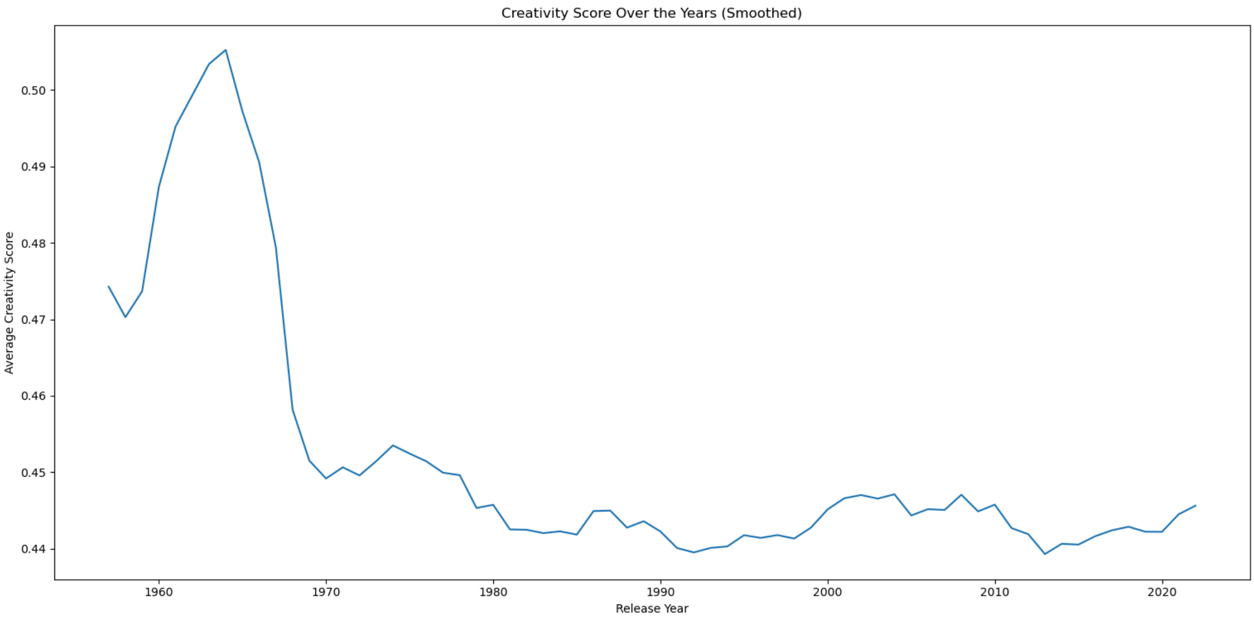
The goal of this analysis is to determine which music styles exhibit the highest levels of creativity based on various lyrical and structural features. Creativity in this context is defined by the complexity and diversity of the lyrics, including the use of unique words, structural elements like bigrams and trigrams, and sentiment measures.



Result:

The visualization shows that **Hip Hop** is the most creative music style. This result suggests that Hip Hop songs tend to have more complex and diverse lyrics, utilizing a wide range of vocabulary, intricate structural elements, and varied sentiments, contributing to a higher overall creativity score.

The creativity over the song's years released



1950s-1960s:

- There is a slight increase in creativity scores in the late 1950s, followed by a significant rise and peak around the mid-1960s, indicating high lyrical creativity.

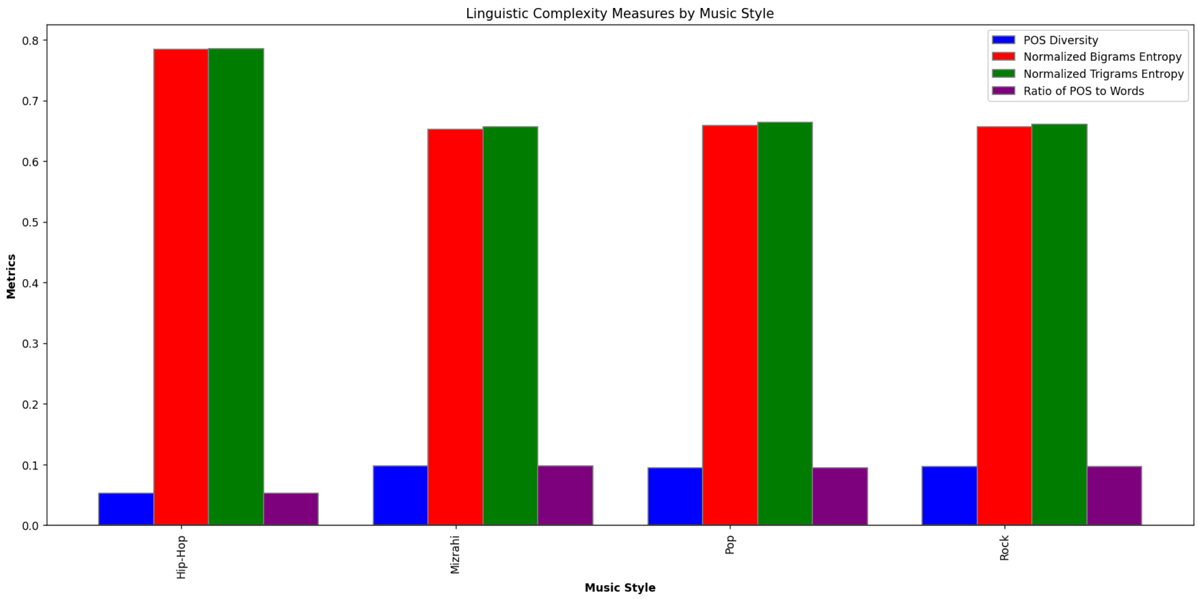
1970s-1980s:

- After the mid-1960s peak, creativity scores decline steeply in the late 1960s and continue to decrease, remaining low throughout the 1970s and 1980s, suggesting reduced lyrical creativity.

1990s-Present:

- The scores stabilize in the 1990s, with minor fluctuations, and show a slight upward trend from the late 2000s onwards, indicating a gradual increase in creativity in recent years.

Analysis and Insights



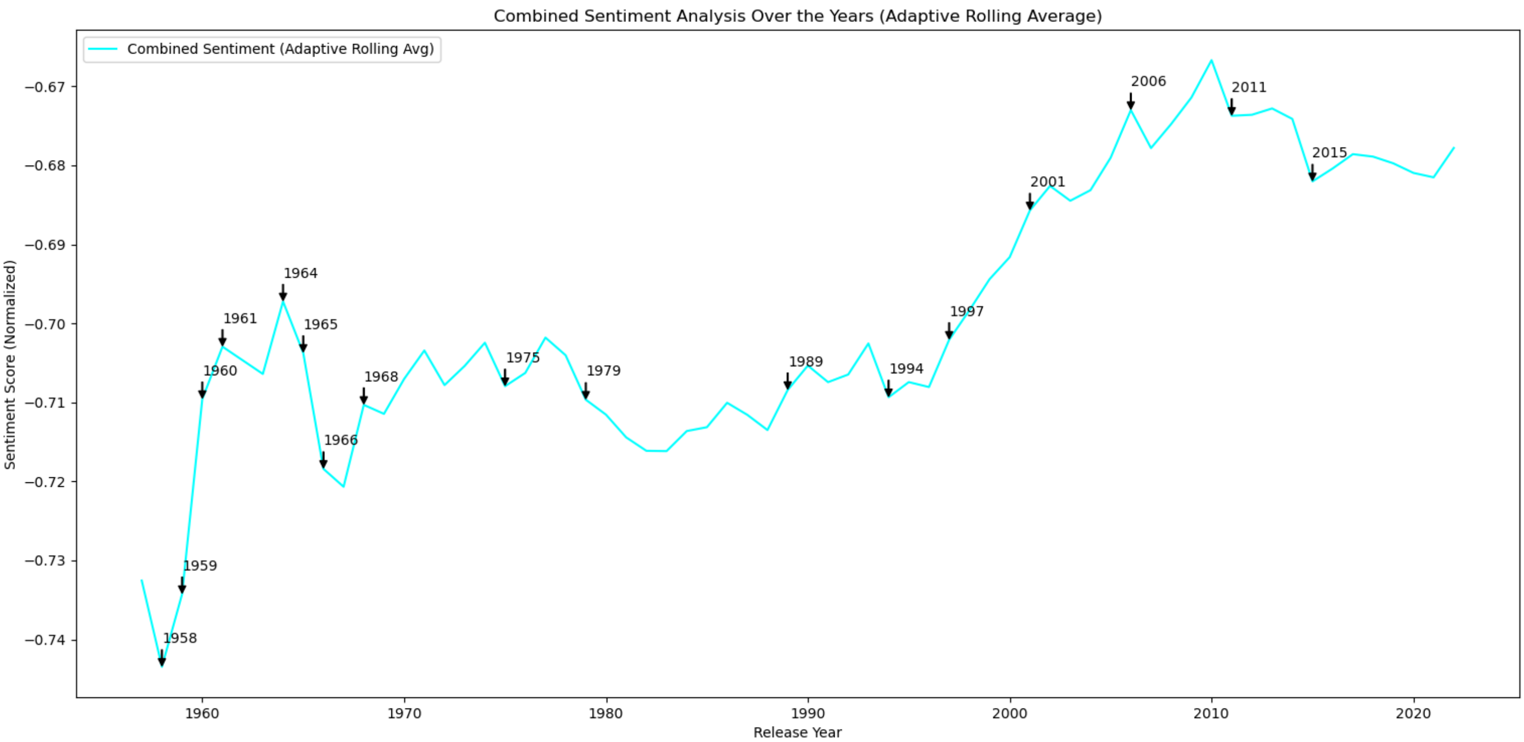
1. POS Diversity:
    - Hip-Hop shows the lowest POS diversity, which might suggest a more consistent use of grammatical structures, often seen in genres that rely heavily on rhythmic and rhyming constraints.
    - Mizrahi, Pop, and Rock display higher POS diversity, indicating a broader range of grammatical constructions that could suggest more complex lyricism or a greater variety of lyric themes.
2. Bigrams and Trigrams Entropy:
    - Hip-Hop has the highest entropy values for both bigrams and trigrams, suggesting a higher degree of unpredictability in word pairings and triplet combinations. This can be indicative of complex lyrical structures which are common in genres that value lyrical dexterity and creativity.
    - Mizrahi, Pop, and Rock have lower entropy values, suggesting more predictability in these genres' lyrical structures. Lower entropy might be indicative of more repetitive or formulaic language use.
3. Ratio of POS to Words:
    - This metric closely aligns with the POS diversity, with Hip-Hop showing the lowest ratio again, indicating fewer types of grammatical structures per word used.
    - The higher ratios for Mizrahi, Pop, and Rock suggest a richer utilization of language forms, which could correlate with more varied and dynamic lyrical content.

Mizrahi, Rock, and Pop genres exhibit similar linguistic patterns, contrasting with the distinct characteristics observed in Hip-Hop. This analysis provides valuable insights into how different genres approach lyricism and linguistic diversity in their music.

Exploring the Influence of Political Events on Song Sentiment

Hypothesis: Political events significantly influence the sentiment expressed in Hebrew song lyrics. Periods of conflict and wars will correlate with more negative sentiments, while times of stability and peace will correlate with more positive sentiments in the lyrics.

israeli political event that we explore according to it.
timeline:https://he.wikipedia.org/wiki/%D7%A6%D7%99%D7%A8_%D7%94%D7%96%D7%9E%D7%9F_%D7%A9%D7%9C_%D7%94%D7%94%D7%99%D7%A1%D7%98%D7%95%D7%A8%D7%99%D7%94_%D7%A9%D7%9C_%D7%99%D7%A9%D7%A8%D7%90%D7%9C



conclusions:

Based on the sentiment analysis chart of Hebrew songs over the years, where higher values on the y-axis indicate more positive sentiment, a brief of the significant years and periods of sentiments:

1. 1958-1966:
   - 1958: The lowest sentiment score observed, potentially influenced by the aftermath of the Suez Crisis in 1956.
   - 1964: Marked improvement in sentiment scores. This period corresponds with the establishment of the Increase in morale Victory - we managed to conquer Sinai

2. 1966-1984:
   - 1967: increase in sentiment after the Six-Day War, military success.
   - 1973: The Yom Kippur War's impact is evident, with declining sentiments in the early 1970s.
   - 1984: Recovery in sentiment scores post-Yom Kippur War, possibly reflecting societal resilience and adaptation.

3. 1985-2000:
   - 1987: The First Intifada leads to a sharp decline in sentiment, indicating the impact of ongoing conflict.
   - 1990-1991: Fluctuations in sentiment during the Gulf War period.

4. 2001-2011:
   - 2001: The Second Intifada brings another sharp decline in sentiment.
   - 2005-2008: Disengagement from Gaza and subsequent military operations like the Second Lebanon War and Operation Cast Lead result in fluctuating sentiments.
   - 2011: Social protests may have contributed to fluctuations but overall reflect a period of political and social change.

5. 2015-2022:
   - 2015: Sentiments remain relatively stable post-2014 Operation Protective Edge.
   - 2022: Slight improvement, possibly reflecting current political and social dynamics, including responses to recent events like Operation Breaking Dawn.

Throughout different periods, significant political events have clearly influenced the sentiment in songs. During times of conflict, such as wars and intifadas, the sentiment in songs tends to be more negative, reflecting the societal mood and the challenges faced by the nation. Conversely, periods marked by peace processes, social change, and economic growth show an increase in positive sentiment, mirroring the hopeful and optimistic outlook of society.

Our analysis supports the hypothesis that political events significantly influence the sentiment expressed in Hebrew song lyrics. The findings highlight a correlation between the societal mood during different political periods and the sentiment reflected in the music. The study underscores the role of music as a powerful medium for expressing and influencing public sentiment, providing valuable insights for understanding the cultural and historical context of musical trends.

It should be noted that there are different factors such as different styles of music that can affect the semantics.