Harmonizing Emotions: Analyzing American Song Lyrics through Sentiment Analysis and Big Data Techniques

Hulya Alpagu[1], Dr. Gulhan Bizel[2]

Saint Peter's University, Data Science Institute, 2641 JFK Boulevard, Jersey City, 07306, NJ, USA

[1]halpagu@saintpeters.edu, [2]gbizel@saintpeters.edu

Abstract:

Music, with its poignant lyrics, is a powerful medium for expressing a range of emotions, connecting listeners to intricate feelings. Our project explores emotional nuances in American song lyrics through sentiment analysis, utilizing advanced NLP techniques like Special Lexicon-Based, VADER, and TextBlob methodologies. The primary goal is to compare these methods and identify the most suitable model, assessing their performance in predicting positive and negative sentiments accurately. Once the most fitting and robust model is identified, our goal is to predict a song's sentiment by analyzing the balance of positive and negative emotions conveyed through its lyrics. Additional efforts are devoted to understanding American song lyrics through big data and time series analysis. The research aims to reveal shifts in emotional tones, distinguishing between positive and negative sentiments over time. Leveraging a diverse dataset, we apply big data techniques for efficient processing and in-depth analysis. Our research contributes extra insights into the American music industry by analyzing positive and negative emotional songs categorized by seasons, tags, and years. This study aims to analyze emotional nuances in American songs, offering insights into the music industry and a unique psychological perspective by discerning emotions through music.

Keywords: *America Song Lyrics, Sentiment Analysis, Special Lexicon-Based, VADER, TextBlob, Big Data*

1. Introduction:

Songs serve as powerful channels for expressing and communicating emotions, deeply influencing the human experience. When songwriters infuse their creations with personal emotions, listeners connect by recognizing echoes of their own experiences in the lyrics. Analyzing these emotions in song lyrics not only reveals broader societal trends but also provides valuable insights for creating works that resonate widely. In essence, exploring emotions through songs forms a potent connection point where individual expression intersects with and reflects broader societal sentiments.

In the research, the goal is to analyze song lyrics, distinguishing between positive and negative emotional tones. Leveraging sentiment analysis, a powerful technological tool rooted in natural language processing (NLP), and insights from big data, the interdisciplinary approach aims to unveil the intricate emotional fabric within songs, enriching an understanding of the connections between music, emotion, and societal trends.

NLP is an artificial intelligence branch dedicated to enabling computers to comprehend, interpret, and generate human language. It encompasses various tasks, such as text tokenization, entity

recognition, part-of-speech tagging, sentiment analysis, machine translation, speech recognition, and text summarization. In our project, we utilize sentiment analysis, a specific aspect of NLP, to extract insights into the emotional nuances within song lyrics, contributing to a deeper understanding of the intricate connections between music, emotion, and societal trends (DeepLearning.AI, 2023).

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique designed to ascertain whether data expresses a positive, negative, or neutral sentiment (Raj, 2023). This method harnesses the power of machine learning, statistics, and NLP to gain insights into the collective thoughts and emotions on a broader scale. By processing written content, sentiment analysis tools unveil the overall positivity or negativity conveyed within the expression, offering a nuanced understanding of public sentiment (GeeksforGeeks, 2023).

Furthermore, sentiment analysis employs diverse methods to discern emotional tones, precisely distinguishing between positive and negative aspects of the sentiment spectrum (Qualtrics, 2023). In our research on American song lyrics, we employ a diverse set of sentiment analysis methods, including Custom Lexicon-Based, Vader, and TextBlob.

Custom Lexicon-based Sentiment Analysis is a method that depends on a foundational sentiment lexicon to discern the emotional tone of the text, categorizing it as positive, negative, or neutral. In this approach, a sentiment lexicon serves as a manually crafted list of lexical features, often words, each labeled based on its semantic orientation as either positive or negative (BOLD Enthusiast, 2023).

TextBlob, a Python natural language processing library, encompasses an efficient sentiment analysis tool, employing a machine learning algorithm trained on labeled text data to classify sentiments and offering polarity scores reflecting the text's degree of positivity or negativity (Es, 2023).

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a sophisticated model designed for text sentiment analysis, demonstrating sensitivity to both the polarity (positive/negative) and intensity (strength) of emotion (Nova & Nova, 2023).

Specialized efforts are dedicated to uncovering the intricacies of American song lyrics through the application of big data and time series analysis. Utilizing a dataset of 33,634 songs, 2,662 artists, and releases spanning from 1963 to 2020, emphasis is placed on the vital role of big data techniques for efficient processing and detailed analysis. In this context, the project aims to leverage big data techniques for more effective processing and comprehensive analysis.

Simultaneously, the deliberate use of time series analysis unveils temporal trends in emotional expressions, offering insights into the evolving nature of sentiments over time. Together, these methodologies enrich our research, providing a holistic understanding of emotional patterns in American song lyrics at both scales and over time.

Big data involves processing and analyzing extremely large and complex datasets that traditional tools cannot handle. It is marked by three key aspects: Volume, representing the sheer size of the

data; Velocity, indicating the speed at which data is generated and processed; and Variety, encompassing the diverse formats of the data (Taylor, 2023).

A time series is a sequence of data points or observations collected, recorded, or measured over successive time intervals. Each data point is associated with a specific timestamp, allowing for the analysis of patterns, trends, and variations over time. Time series data is commonly used in various fields for tasks such as forecasting, trend analysis, and understanding temporal behaviors (Wikipedia contributors, 2024).

2. Background and Related Work:

The objective of this exhaustive literature review is to thoroughly investigate the spectrum of positive and negative emotions within American song lyrics, with a particular emphasis on employing advanced techniques such as sentiment analytics, time series analysis, and big data methodologies.

In the domain of music emotion analysis, this study focuses on automating the process through the utilization of IoT and LSTM networks. The objective is to enhance users' intuitive comprehension of emotional nuances in music. Notably, the LSTM-based model, integrated with the Sequence-to-Sequence (STS) framework, has exhibited superior performance compared to conventional methods. These findings not only contribute to the current understanding of musical emotion analysis but also offer valuable insights for future research and applications in this field (Cao & Park, 2023).

This paper surveys music emotion recognition studies, guiding machine learning implementation (EEG, CCN, IADS). Despite current model strengths, limitations persist, highlighting the need for crucial future work in advancing sophisticated systems (Li, 2023).

In this NLP study, Filipino sentiments during crises on Twitter are examined. Unveiling a dominant neutral tone, followed by positive and negative expressions, the research delves into emotional patterns specific to crises like earthquakes, inflation, and typhoons. Notably, the pandemic introduces a distinctive sequence of neutral, negative, and positive tones. This research offers valuable insights into how Filipinos articulate emotions on social media amidst challenging situations (Villasor & Baradillo, 2024).

This research employs TextBlob and VADER analyzer on historical tweets, exploring emotional responses during Nigeria's COVID-19 pandemic. Emphasizing social media's potential for informed decision-making, the study addresses challenges like misinformation and reveals diverse public perceptions of COVID-19, from viewing it as a persistent threat to optimism about eventual triumph (Abiola et al., 2023).

This study addresses challenges faced by Indonesian MSMEs, highlighting the role of marketplaces in boosting competitiveness. Analyzing six platforms, Blibli.com emerges as the top choice for MSME marketing using Lexicon-based and naïve Bayes methods. The research contributes scientifically and practically, optimizing digital marketing strategies for MSMEs. Sentiment analysis, based on Twitter reviews, emphasizes the importance of balanced data for

accuracy. Recommendations include feature enhancement, manual labeling, and exploration of alternative machine-learning techniques for improved results (Hoiriyah et al., 2023).

The research unveils a real-time Twitter sentiment prediction system for Moroccan universities, integrating big data analytics and sentiment analysis. Employing Twitter Streaming API, Apache Kafka, Apache Spark, Elasticsearch, and Kibana, the system comprises offline sentiment analysis and a real-time prediction phase. The Random Forest classifier, trained on historical French tweets, achieves remarkable accuracy. Findings from the online phase provide valuable insights for higher education decision-making (Lasri et al., 2023).

The study investigates sentiment changes on Twitter for eight restaurant chains before and after the US federal menu labeling law in 2018. VADER sentiment analysis and Controlled Interrupted Time Series (CITS) reveal minimal shifts, emphasizing the need for sustained efforts to influence sentiment toward unhealthy chains. Post-legislation, the study emphasizes the need for increased attention from organizations to amplify positive effects through social media (Hswen et al., 2023).

As a result of our literature review, a comprehensive understanding of the methodologies and approaches used in the analysis of emotions in song lyrics forms the basis of our study. The integration of advanced sentiment analysis technologies and insights from studies on big data and time series analysis contributes to the depth and breadth of our research.

3. Methodology:

3.1 Data Collection

The dataset for this research was sourced from Genius (Partridge, 2024) between September 2019 and January 2020, integrating information from three distinct files: "lyrics", "Song Info" and "Artist Info". The "lyrics" file, formatted in the Julia scripting language, comprises a comprehensive collection of 33,634 songs along with their respective lyrics. In the " Artist Info " file, presented in JSON format, details for 2,662 artists are available, including their "URL name," the number of followers on their Genius artist page, and an exhaustive list of songs to which they have contributed. Additionally, the ' Song Info ' file contains data on songs released between 1963 and 2020, presented in JSON format. This includes the “URL name”, primary artist, page views, the number of contributing users to the song, and genre tags.

As illustrated in Figure 1, initially, a new column titled 'season' was added to the 'Song Info,' discerning the season of song releases based on the 'Release Date' column. Following this, the dataset was constructed by concatenating the 'Song Info' and 'Artist Info' files using 'URL name' as the key. Subsequently, this merged file seamlessly integrated with the 'Song' column from the ' Lyrics' file, forming a structured data frame poised for in-depth analysis.

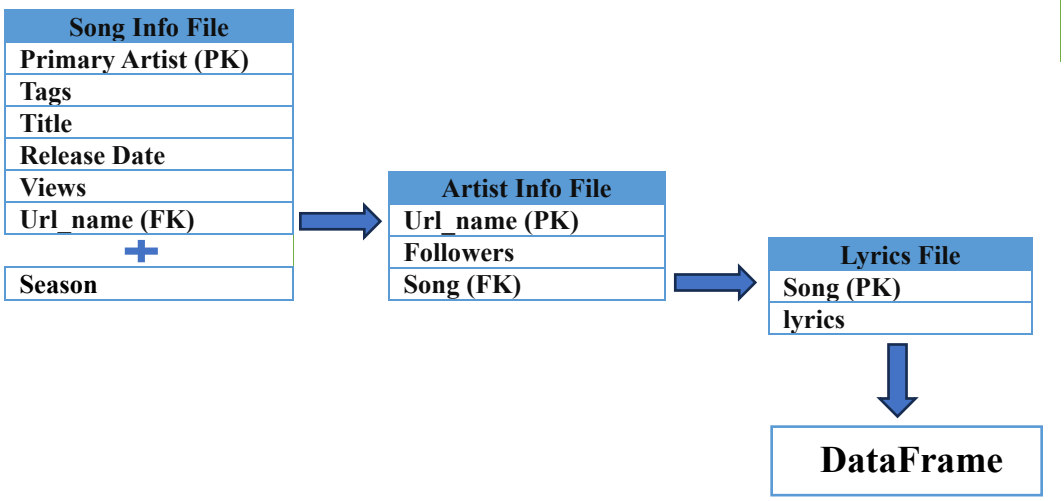

Fig 1: Formation of the Dataset through the Merging of Three Files.

3.2 Sentiment Analysis

In the project, Custom Lexicon-Based, VADER, and TextBlob sentiment analysis methods were utilized for their proven effectiveness. Our primary aim is to meticulously compare these approaches to identify the most suitable model for seamless integration. The selection criteria involve assessing each model's performance in accurately predicting both positive and negative sentiments within song lyrics. Once the most fitting and robust model is determined, our objective is to predict a song's sentiment by analyzing the balance of positive and negative emotions conveyed through its lyrics. For example, if a song exhibits a higher prevalence of negative words compared to positive ones, its sentiment will be categorized as negative.

3.2.1 Custom Lexicon-Based Sentiment Analysis

For the America Song Lyrics project, we developed a sentiment analysis technique that uses a specialized lexicon to identify distinctive sentiment expressions. This method enhances precision by aligning closely with linguistic nuances, providing a nuanced understanding of sentiment.

3.2.2 Vader Sentiment Analysis

We leveraged VADER's capabilities, tapping into its proficiency in processing informal language and employing sophisticated valence-aware scoring. The integration of its pre-trained model effectively captured intricate emotions in American song lyrics.

3.2.3 TextBlob Sentiment Analysis

We harnessed TextBlob's robust sentiment analysis capabilities to explore emotional nuances within American song lyrics, enriching our project with valuable insights.

3.3 Time Series Analysis

To unravel the distribution of positive and negative sentiments in American song lyrics, we embraced a methodologically robust approach, harnessing the power of time series analysis. This thorough exploration delved into the emotional dynamics of American song lyrics, uncovering a nuanced distribution of positive and negative sentiments correlated with the respective release dates of the songs.

4 Analysis:

Figure 2 displays a word cloud featuring the most frequently used words in American song lyrics. This visualization provides insights into the prevalence of specific terms in song lyrics, offering valuable information about common word usage.

Fig. 2: Word Cloud Analysis of American Song Lyrics

Figure 3 visually represents usage distributions by spotlighting the 15 most common words in American song lyrics. The visualization indicates that the frequently occurring words in the dataset tend to convey a negative sentiment.

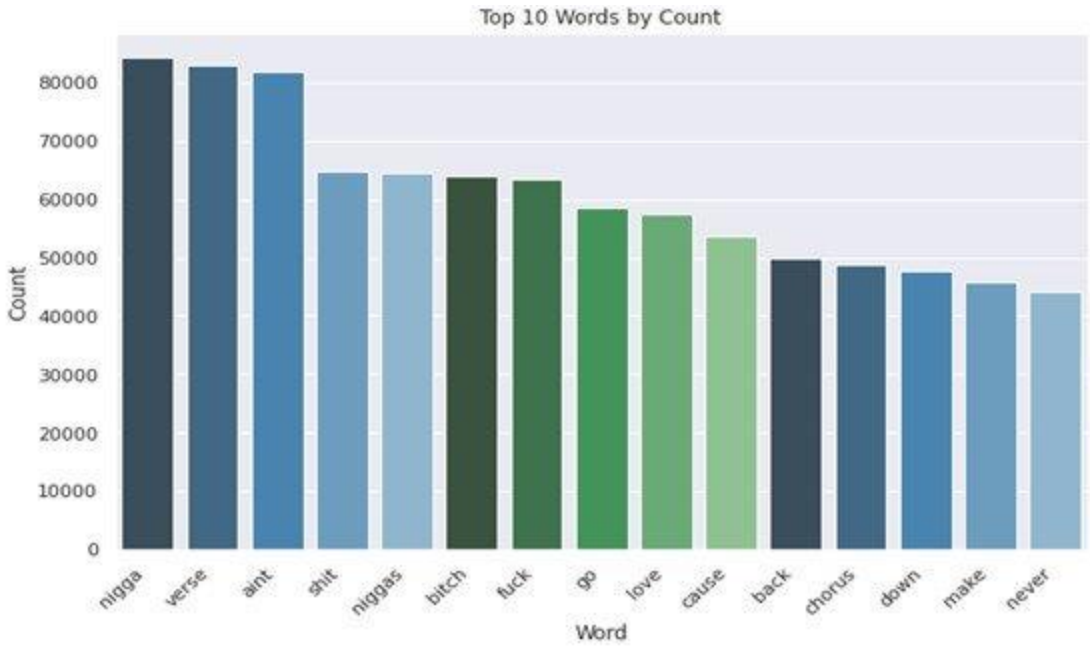

Fig. 3: Exploring Word Frequencies: Top 15 Words in American Song Lyrics

In Figure 4, the percentage distribution of various music genres in the market is visually represented. Upon closer scrutiny of Figure 4, it becomes apparent that Rap songs dominate the market with the highest presence, followed by Pop and Mix songs. This observation underscores the substantial popularity of Rap songs in the American music market. While Pop and Mix songs may not reach the same level of popularity as Rap, they are closely followed and enjoy widespread listenership.

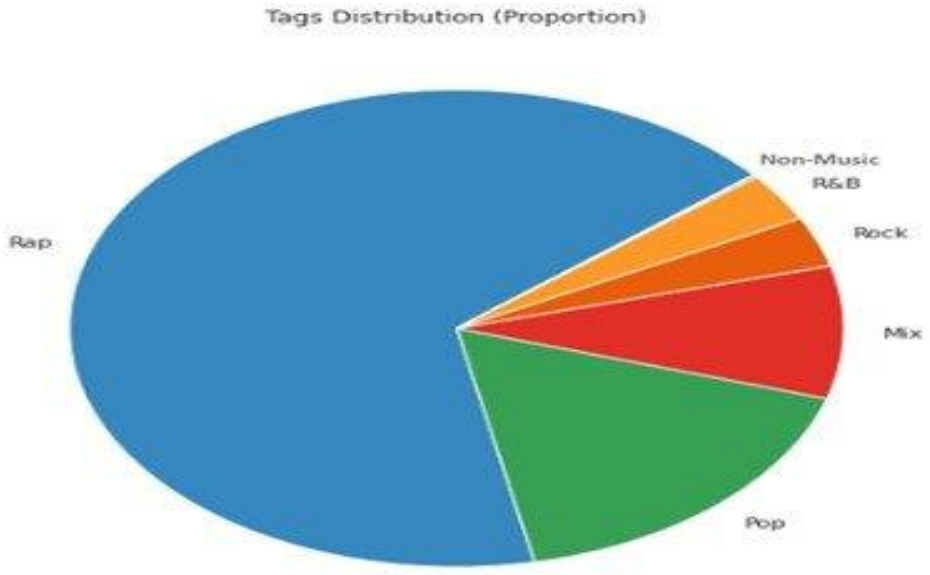

Fig. 4: Genre Distribution in the American Music Market

Figure 5 provides a detailed perspective on the presence of songs in the market throughout the seasons. Upon closer examination, it is evident that songs in the fall season have the highest market presence, followed by the summer season. Additionally, songs in the winter season exhibit the lowest market presence.

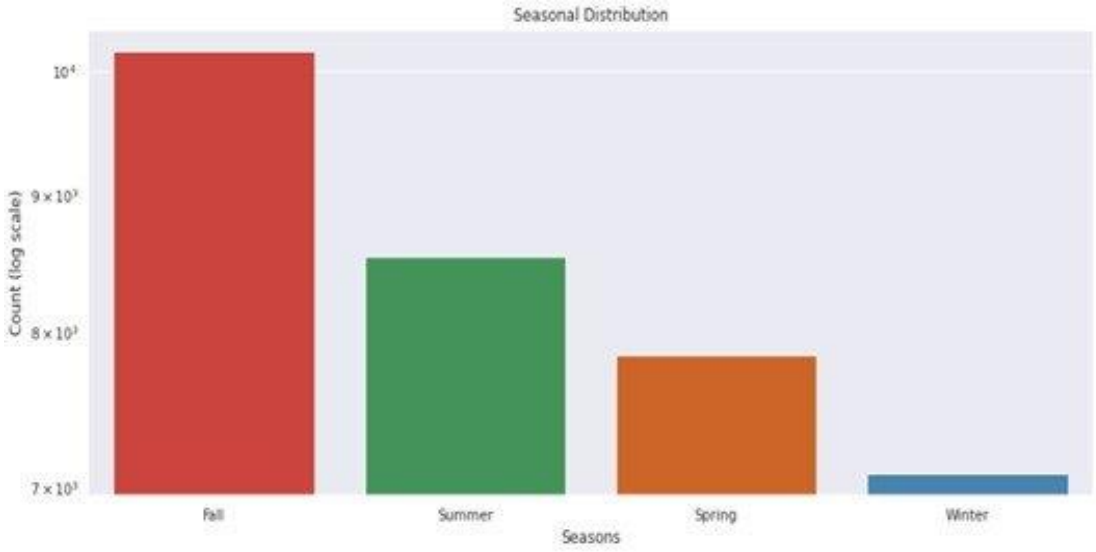

Fig. 5: Seasonal Presence of Songs in the Market

5 Result:

After comparing sentiment analysis methods, including Custom Dictionary Based, VADER, and TextBlob, as illustrated in Figure 6, it became evident that the Vader method outperformed others in predicting both positive and negative sentiments for our project. This outcome led us to adopt the Vader method as the preferred model, and we proceeded with our project using this approach.

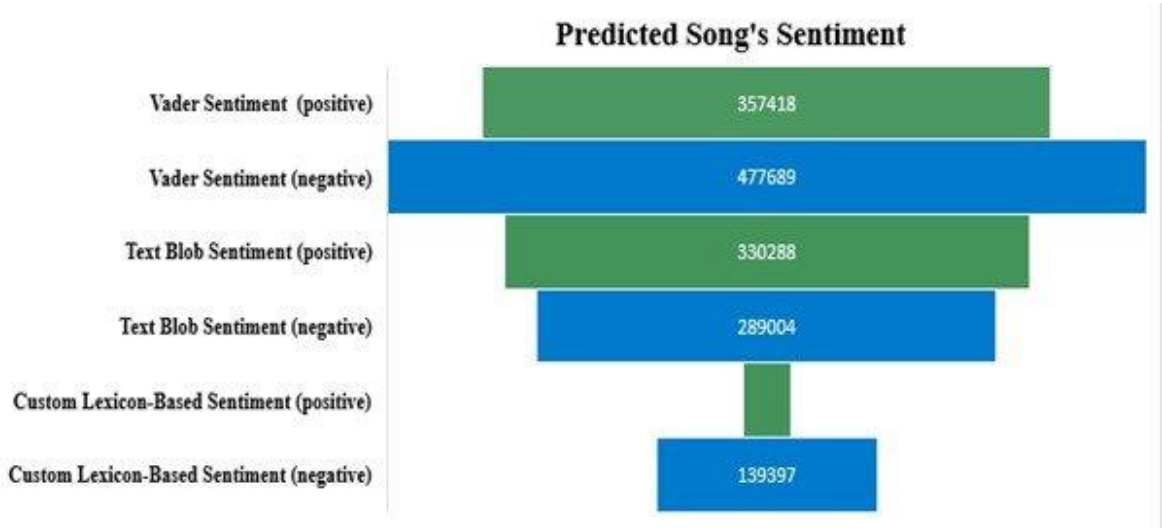

Fig. 6: Optimizing Sentiment Analysis: Choosing VADER as the Preferred Model

Figure 7 illustrates the distribution of positive and negative emotions within the dataset of American song lyrics. Notably, the visualization reveals a predominance of negative emotions compared to positive ones. This observation underscores that the overall sentiment conveyed by the songs tends to lean predominantly toward a negative emotional tone.

final_sentiment_vader	count
positive	10361
negative	23273

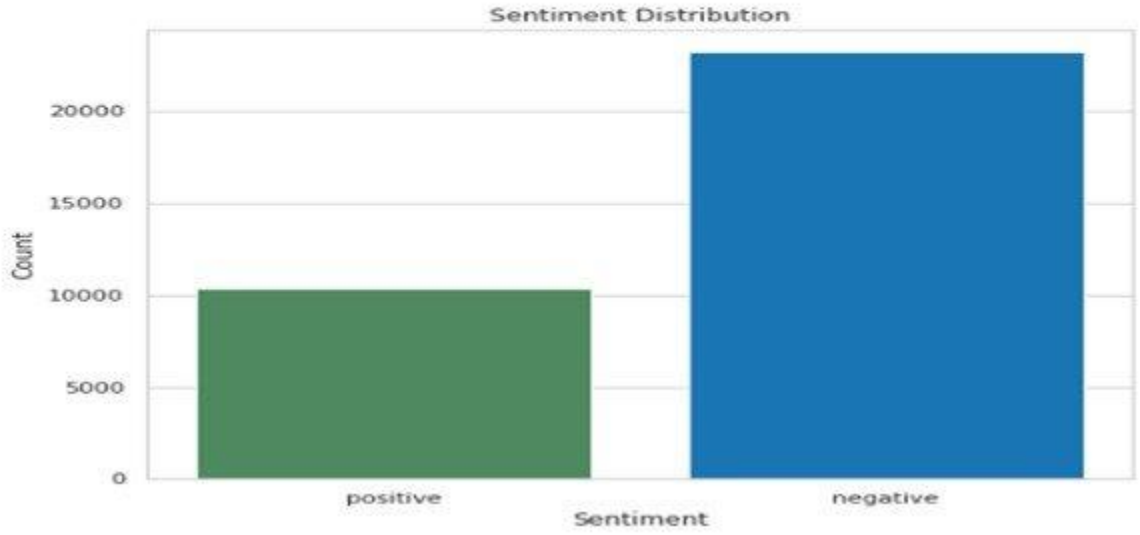

Fig. 7: Exploring Sentimental Tones in American Song Lyrics: An Analysis of Positive and Negative Emotion Distribution

Figure 8 visually represents the distribution of the interaction between positive and negative emotions in the market for various music genres. When looking at Figure 8, it is evident that negative emotional rap songs dominate the market presence, followed sequentially by positive emotional rap, positive emotional pop, and negative emotional pop songs. The analysis section, informed by Figure 4, establishes the overarching impact of rap songs on shaping the market landscape. Further scrutiny of Figure 8 reveals that most of these rap songs achieve popularity due to their negative emotional content. Figure 8 illustrates that the findings presented in Figure 7 elucidate why the quantity of negative emotional songs in the market is notably higher than that of positive emotional songs.

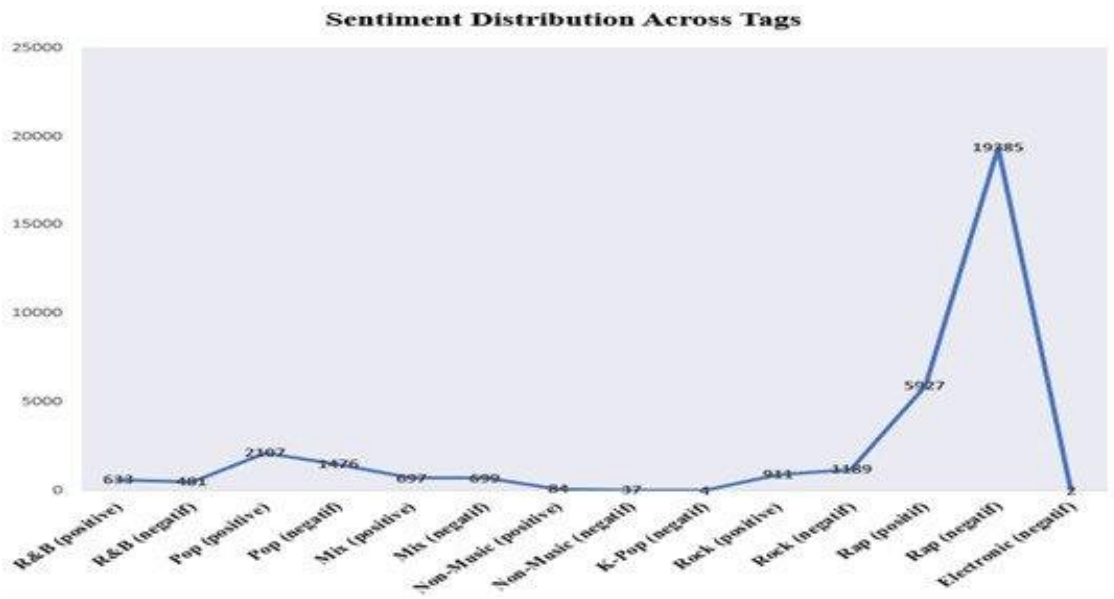

Fig. 8: Analyzing the Emotional Distribution of Songs Across the Market

The chart in Figure 9 illustrates whether the 15 most viewed American songs are characterized by positive or negative emotions. Upon examining Figure 9, it becomes evident that only 4 out of the 15 songs reflect positive emotions. This graph highlights a tendency among individuals to show greater interest in songs with negative emotions and less interest in positive ones.

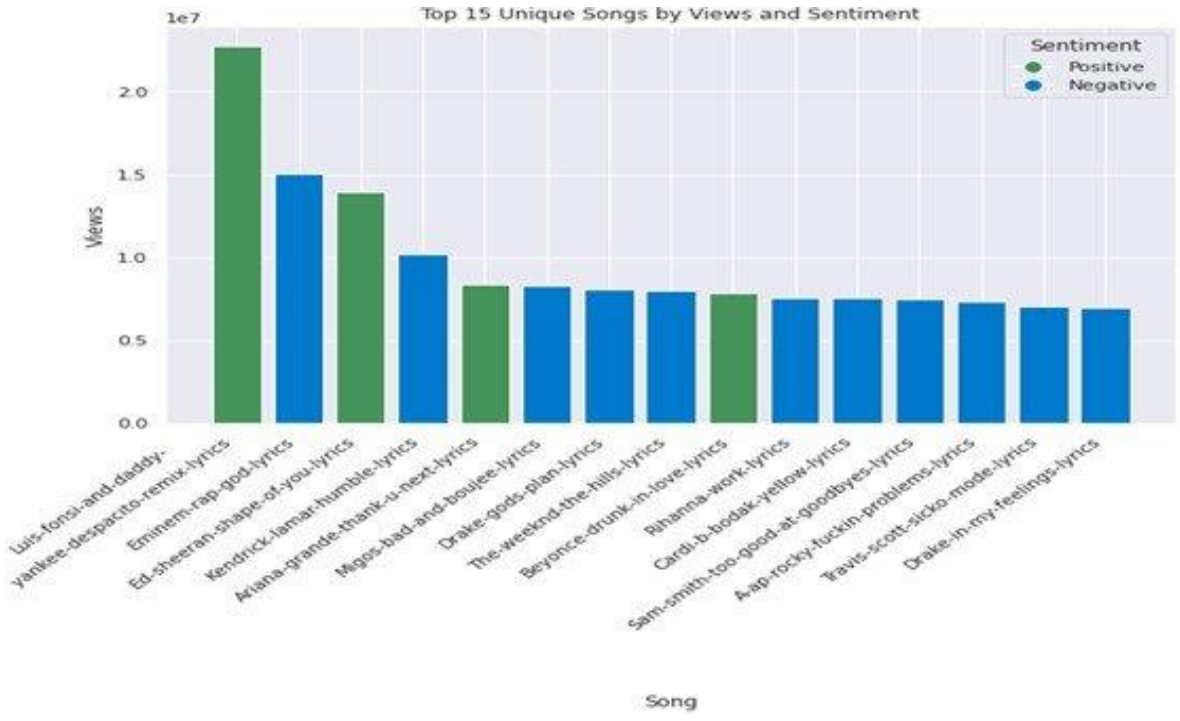

Fig. 9: Emotional Trends in the Top 15 Most Viewed American Songs: A Focus on Positive and Negative Expressions

Figure 10 offers a detailed analysis of the interplay between positive and negative emotions in American song lyrics, categorized by seasons. Notably, each season portrayed in Figure 10 reveals a distinct prevalence of songs that evoke negative emotions compared to those expressing positivity. What adds an intriguing dimension to this observation is the temporal pattern, where the majority of songs with negative themes are released after the winter season, particularly during the summer season. This unique temporal trend, as illustrated in Figure 10, underscores a notable inclination towards songs with negative emotions, even in the traditionally vibrant atmosphere of the summer season.

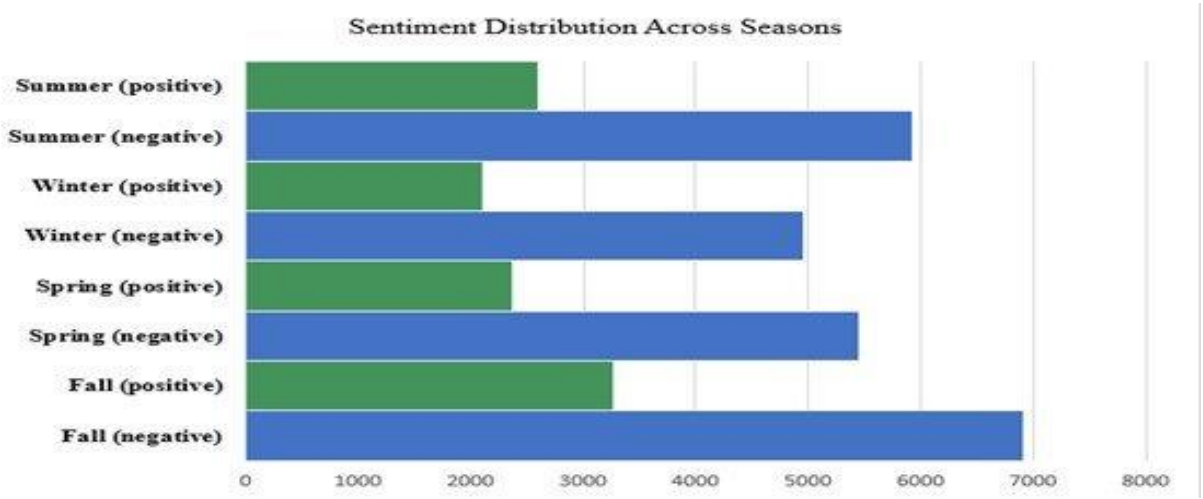

Fig. 10: Seasonal Trends in Song Presence and Sentiment Distribution: A Detailed Analysis

Figure 11 unveils a comprehensive time-series analysis, elucidating the nuanced distribution of positive and negative emotions in American song lyrics according to their release dates. As depicted in Figure 11, there is a noticeable surge in both positive and negative emotions in American song lyrics after 1995, reaching a peak around 2010 to 2020. Furthermore, the time-series analysis indicates that after 1995, lyrics with negative sentiment became more prevalent than those with positive sentiment.

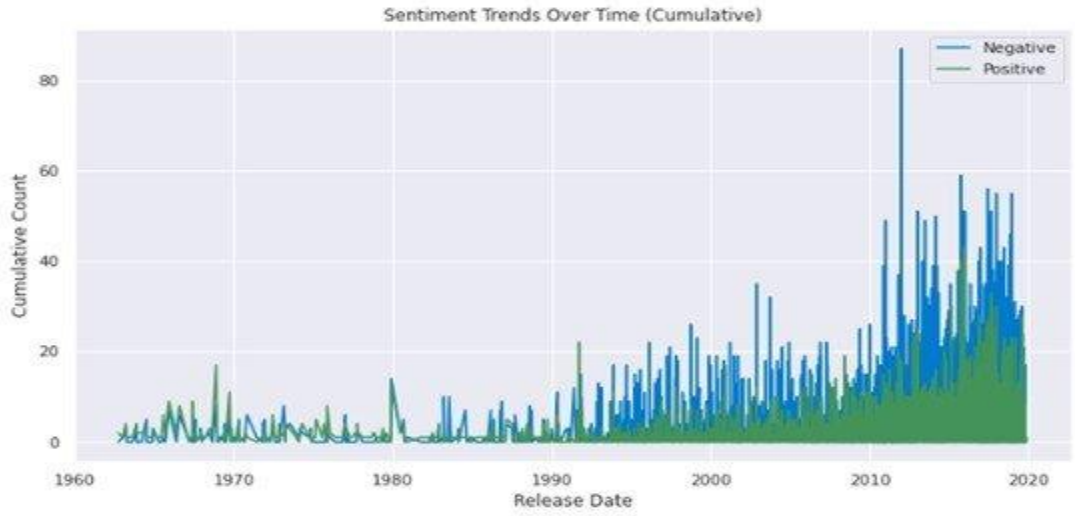

Fig. 11: Exploring the Evolution of Emotional Content in American Song Lyrics: A Time-Series Analysis

6 Conclusion:

Through the efforts invested in our project, a notable surplus of songs infused with negative emotions has been observed in the market compared to their positive counterparts, emphasizing a prevailing preference for songs with negative emotional content among the audience. Furthermore, our project delivers invaluable insights poised to enhance the American music industry. For example, it provides detailed analyses categorized by seasons, tags, and years, offering comprehensive information regarding the prevalence of both positive and negative emotional songs in the market. This study serves as an essential guide for professionals seeking to grasp the intricacies of the American music industry, empowering them to make informed and strategic decisions.

7 Future Work:

In the continuation of this study, there is a need for a more in-depth psychological investigation to understand why negative emotional songs dominate the market compared to positive ones and why they are preferred by listeners. Additionally, to better grasp the dynamics of the music industry and adapt to changing music trends, more data and analysis will be necessary to identify which emotional genres may take prominence in future projects. This could assist music producers, artists, and marketers in making strategic decisions.

Reference:

Ken, P. (2024, January 12). Genius | Song Lyrics & Knowledge. https://genius.com/

DeepLearning.AI. (2023, January 11). Natural Language Processing (NLP): A Complete Guide. https://www.deeplearning.ai/resources/natural-language-processing/

Raj, N. (2023, November 6). Starter's Guide to Sentiment Analysis Using Natural Language Processing. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/nlp-sentiment-analysis/

GeeksforGeeks. (2023, December 21). What is Sentiment Analysis?

https://www.geeksforgeeks.org/what-is-sentiment-analysis/

Qualtrics. (2023, September 28). What is sentiment analysis and how can users leverage

Qualtrics? https://www.qualtrics.com/experience-management/research/sentiment-analysis/

BOLDEnthusiast. (2023, September 11). Sentiment Analysis – The Lexicon-Based Approach. Microsoft Dynamics and NetSuite Partner & Dynamics CRM Consultants in San Diego. *https://www.alphabold.com/sentiment-analysis-the-lexicon-based-approach/*

Es, S. (2023, August 30). An In-depth Comparison of Sentiment Analysis Tools in Python: Evaluating TextBlob, Vader Sentiment, Flair, and Custom Implementation. *https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair*

Nova, & Nova. (2023, April 23). Getting Started with Sentiment Analysis using VADER. AITechTrend - Further into the Future. *https://aitechtrend.com/getting-started-with-sentiment-analysis-using-vader/*

Taylor, D. (2023, December 30). What is Big Data? Introduction, Types, Characteristics, Examples. Guru99. *https://www.guru99.com/what-is-big-data.html*

Wikipedia contributors. (2024, January 6). Time Series. Wikipedia. *https://en.wikipedia.org/wiki/Time_series*

Cao, & Park. (2023, December 12). The Analysis of Music Emotion and Visualization: Fusing Long Short-Term Memory Networks under the Internet of Things, 11, 141192 - 141204. *https://ieeexplore.ieee.org/abstract/document/10354332*

Li, Z. (2023, April 1). Emotion Recognition of Music Based on Machine Learning Scenarios. Highlights in Science Engineering and Technology, 39, 144–150. *https://doi.org/10.54097/hset.v39i.6515*

Villasor, H. D. B., & Baradillo, D. G. (2024, January 6). Natural Language Processing Employing Sentiment Analysis on the Public Voice of Filipinos During Crisis Situations. EPRA International Journal of Multidisciplinary Research (IJMR), 10(1), 80–89. *https://eprajournals.net/index.php/IJMR/article/view/3487*

Abiola, O., Abayomi-Alli, A., Tale, O. A., Misra, S., & Abayomi-Alli, O. (2023). Sentiment Analysis of COVID-19 Tweets from Selected Hashtags in Nigeria Using VADER and Text Blob Analyzer. Journal of Electrical Systems and Information Technology, 10(1). *https://doi.org/10.1186/s43067-023-00070-9*

Hoiriyah, H., Qomariya, N., Darmawan, A. K., Walid, M., & Efenie, Y. (2023, September). "Sentiment Analysis on LGBT Issues in Indonesia with Lexicon-Based and Support Vector Machine Algorithms." Jurnal Pilar Nusa Mandiri, 19(1), 27–36. *https://doi.org/10.33480/pilar.v19i1.4183*

Lasri, I., Riadsolh, A., & Elbelkacemi, M. (2023, March 1). "Real-time Twitter Sentiment Analysis for Moroccan Universities using Machine Learning and Big Data Technologies." International Journal of Emerging Technologies in Learning (Ijet), 18(05), 42–61. *https://doi.org/10.3991/ijet.v18i05.35959*

Hswen, Y., Moran, A. J., Von Ash, T., Prasad, S., Martheswaran, T., Simon, D., Cleveland, L. P., Brownstein, J. S., & Block, J. P. (2023, December). "The impact of the federal menu labeling law on the sentiment of Twitter discussions about restaurants and food retailers: An interrupted time series analysis." Preventive Medicine Reports, 36, 102478. *https://doi.org/10.1016/j.pmedr.2023.102478*