# THE INFLUENCE OF GEO-LOCATION MENTIONS IN THE PROJECTION OF A MUSIC

**Olamide Shogbamu[1], Olubayo Adekanmibi[2] & Anthony Soronnadi[3]**
Data Science Nigeria (DSN), Lagos, Nigeria
{shogbamu,olubayo,anthony}@datasciencenigeria.ai

## ABSTRACT

Music is a multifaceted socio-cultural phenomenon characterized by diverse genres that have evolved in specific geographical regions. This paper delves into Nigeria's relatively unexplored intersection of geography and music. While previous research has addressed music geography, particularly in the context of music's origin, evolution, and diffusion, the role of geographical language and lyrical elements in Nigerian music still needs to be studied. This study focuses on bridging this gap. Utilizing a Geo-Contextual Semantics approach, we conducted a thorough analysis of Nigerian music content to investigate the impact of geolocation on music success. Our study uncovers a noteworthy relationship, contingent upon other factors, between mentions of geographic locations and music success potential in Nigeria.. This research does not seek to limit the creative process of music creation to a rigid framework. Instead, it suggests establishing a framework that can enhance the growth of music. The strategic application of this framework has the potential to illuminate uncharted territories in the Nigerian music landscape and foster boundaryless collaboration among artists, thus contributing to the enrichment of the music industry.

**Keywords:** Geo Semantic, Natural Language Processing,
Music, Lyrics, Location mention

## 1 INTRODUCTION

A society's uniqueness is primarily determined by how it uses language. Its inventive creation and expression of proverbs, idioms, word plays, poetry, and music lyrics convey its beauty. Lyrics are an essential form of language and communication. They are an excellent resource for learning a language. They can help you improve your vocabulary, listening, pronunciation, and comprehension skills.

Geolocation mention is a term used to capture the geographic location mentioned in the lyrics of a song. It is the instance where a song's lyrics explicitly reference specific geographic locations. This could include mentions of cities, countries, landmarks, or other places identifiable on a map.

This study investigates the influence of geolocation mentioned in Nigerian secular music lyrics. The dataset is lyrics released between 2022 and 2023. We employed GeoNLP as a novel analytical tool. It is a combination of Geo-semantic analysis and Natural Language Processing. We hypothesize that the frequency and thematic content of geographic references within lyrics reflect underlying societal trends and, potentially, music success. We are utilizing a corpus of 100 songs from mostly Afrobeat's genre. The introduction and the innovative application of GeoNLP technology allow us to analyze and comprehend the spatial dimensions of lyrical content.

Through this extensive analysis of lyrics, we seek to identify patterns, trends, and meaningful conclusions, shedding light on the underlying dynamics of the connection between geolocation mention and music success. This study offers a new perspective on music research. It also shows how GeoNLP can uncover cultural insights through language analysis, demonstrating its interdisciplinary potential.

## 2 Literature Review

The study of music lyrics has been the topic of much scholarly investigation. Varnum et al. (2021) investigated whether music lyrics were becoming increasingly simple overtime. He tested the assumption that as lyrical simplicity increases, so does the number of innovative song choices. Werth (2021) decided to do a music-matching experiment using natural language processing. He wanted to replicate the unsyncing rhythm in the music recommender app. He did this by using song lyrics.

Leyshon et al. (1998) define the topic as studies of music and cultural politics amid the rise to prominence of "new cultural geography," Since then, researchers have become increasingly interested in the relationship of geography and music, recognizing the enormous impact that place has on musical tastes, economic inequality, and political identity. Johansson & Bell (2009) work investigates the breadth and diversity of popular music geographies. Pinch & Bijsterveld (2012) and Sterne (2012) guide the still-developing interdisciplinary discipline of sound studies, including topics such as environmental sound, sound art, scientific data, noise pollution, and the sociology of technology. Revill (2017) thoroughly reviewed the thread of musical geographies.

Altaweel (2020) work 'The Geography of Music' confirmed a clear link between geography and music. He also asserts that many different forms of music frequently incorporate location or geography into their lyrics. Bengier (2020) released a list of the 50 best songs with a location in the title through Bestlife, a news and entertainment platform. Boughanmi & Ansari (2021) developed a machine-learning framework that predicts the popularity of music albums by analyzing acoustic parameters like energy levels, liveness, and dance.

Pastukhov (2022) and Merritt et al. (2023) worked on predicting music success using AI. An argument has been made that music and geographic location are linked. People have made instances of music success prediction. However, the analysis of the geographic location mentioned in the music lyrics needs to be improved. The probability of music success based on geolocation mentions has yet to be implemented. This study seeks to bridge this gap by presenting a novel GeoNLP approach for analyzing and understanding geographic location mentions in music lyrics. Based on geolocation mentions, it also investigates if we can predict song popularity.

## 3 Methodology

This research combines exploratory and analytical methods to study the influence of geospatial references in 2023 song lyrics on music success. By focusing solely on songs from 2023, temporal bias is avoided, enabling the exploration of correlations between location mentions and success metrics. The methodology includes identifying and analyzing geospatial features in lyrics, utilizing manual coding, natural language processing (Named Entity Recognition), and Geospatial techniques. Comparative analysis will be conducted between songs with and without location mentions, while ethnographic may add depth to the location mentions.

### 3.1 Data Collection

The analysis chose 163 Spotify songs based on 2023 social trends. Release date, popularity, explicit material, and regional relevance were among the metrics gathered. Lyrics were manually obtained from AZLyrics, Lyrics (a), and Lyrics (b) to ensure completeness and prevent encoding errors. Spotify was chosen as the primary data source due to its status as the most popular audio streaming service, boasting over 602 million users, including 236 million subscribers.

Futhermore, TikTok video counts were acquired from SongStats and manually validated from TikTok. TikTok was chosen as an additional data source because of its efficiency and widespread popularity as a medium for spreading content and showcasing talent. With over 4.5 billion downloads worldwide and availability in more than 150 markets, including Nigeria, TikTok has become one of the most popular apps globally, appealing to users across all age groups.

Data wrangling aggregated the information into a single dataset, managing NaN values and resolving discrepancies, yielding a final balanced dataset.

### 3.2 Feature Extraction and Feature Engineering

New columns were created from the lyrics, using Named Entity Recognition to detect mentions of specific locales. A count column was introduced to monitor the frequency of location mentions. As a result, a binary column was added to indicate if a lyric has a place mention (1) or not (0).

We used feature engineering to prepare the dataset for machine learning model training. The target variable was defined as the weighted average of popularity % and TikTok video count, which was then normalized using the min-max normalization technique. The count column was also normalized using the same approach.

### 3.3 GeoSpatial EDA

#### 3.3.1 Toponym resolution

The location names were combined into a single list to ease consecutive geocoding. We then used the Nominatim geocoder from Open Street Map and the geopy library to determine the locations' geographical coordinates (latitude and longitude) in the consolidated list. Then, a count of unique location names was performed, and the resulting frequency distribution was analyzed and graphically depicted. A spatial join approach was used to represent this distribution on a map.

### 3.4 Training and Prediction

This study employs different machine learning models to predict the success of songs, which are classified as successful (1) or average (0). The models used are Support Vector Machines (SVM), K Nearest Neighbors (KNN), Random Forest (RF), and Extreme Gradient Boosting (XGBoost).

- *Support Vector Machine (SVM)* is a supervised learning model designed to discover the optimal hyperplane for categorizing data points or predicting their values.
- *K-Nearest Neighbour (KNN)* is a machine learning method that makes predictions by selecting the K data points that are closest to a new data point and using their labels. It works on the notion that similar data points provide similar results. We chose K (the number of neighbors to consider) to 3 because our theory is that data points with location mentions can help forecast similar values for inputs with similar location mentions.
- *Random Forest (RF)* is a supervised classification approach that uses bagging to combine many decision trees and average their probabilistic predictions, rather than allowing each tree to vote for a single class. In our case, we set the randomness of splitting and sampling (random state) to 42 and use 50 decision models.
- *Extreme Gradient Boost (XGBoost)* is a machine learning technique that uses multiple decision trees to make predictions in classification and regression tasks. It integrates a variety of strategies to improve the gradient boosting algorithm's performance and speed, such as regularization, parallelization, tree pruning, and missing value handling.

### 3.5 Features and Assumption

We introduced a column called 'count' that shows the frequency of geographical mentions in song lyrics. In addition, we added a feature called 'explicit' that we used to indicate whether the song is trench or lush. These variables, combined with a binary indication for location mention, are used as inputs to our predictive algorithms.

Our objective variable, 'weighted_average', is calculated as the sum of Spotify popularity and TikTok usage, normalized to a range of 0 to 1. We assigned a weight of 70% to Spotify popularity and 30% to TikTok usage. This is because Spotify provides comprehensive data on songs, including metrics such as play counts, and popularity. This depth of data allows for more nuanced analysis and temporal aggregate compared to TikTok, which primarily offers video view counts and engagement metrics. We then divided songs into two categories: successful (values of 0.5 to 1.0) and normal/average (values of 0.0 to 0.4999).

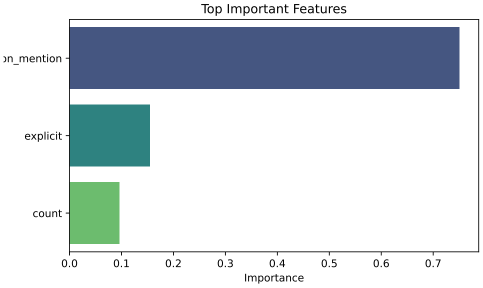

Figure 1: Top Important Features

A study of feature importance found that the count of location mentioned was the most significant, followed by explicit and location mention.

# 4 Discovery

From the dataset utilized in this investigation, significant insights have been discovered that hold the potential for informing decisions or providing justifications.

## 4.1 Distribution and Frequency of Location Mentions

Within a corpus of 70 songs with location references, 152 mentions come from 30 locations. These references are widely dispersed globally, with representation spanning all cardinal directions. Interestingly, despite the corpus being primarily composed of Nigerian music, Ghana has the highest frequency of place citations, followed by London and Lagos. Notably, the latter contains two references, labeled as 'Lagos' and 'Eko,' which both refer to the same geographical place. Consolidating these references would place Lagos ahead of Nigeria but behind Ghana in frequency. Other prominent destinations include Africa, Lekki, Las Vegas, and Colorado, among many more.

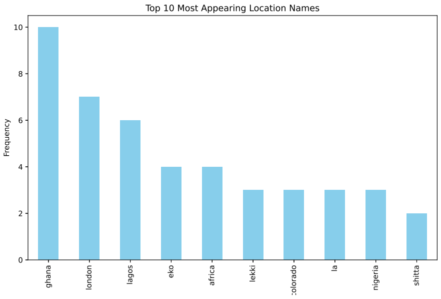

Figure 2: Top 10 Most Appearing Location Names

Furthermore, within the overall significance of Lagos as a frequently mentioned location, Lekki receives the most mentions, followed by Gbagada. The mention of Nigeria and Lagos respectively in the data below is represented using the centroid method.

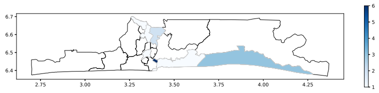

Figure 3: Geographical Distribution of Location Counts accross Lagos

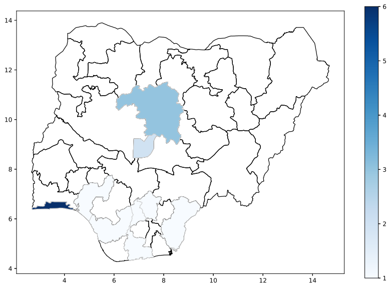

Figure 4: Geographical Distribution of Location Counts accross Nigeria

### 4.2 Correlation coefficient

A direct linear regression analysis reveals that most of the relationship between the variables appears to be stochastic, despite the presence of several data points exhibiting a weak positive correlation and only a few instances demonstrating a weak negative correlation.

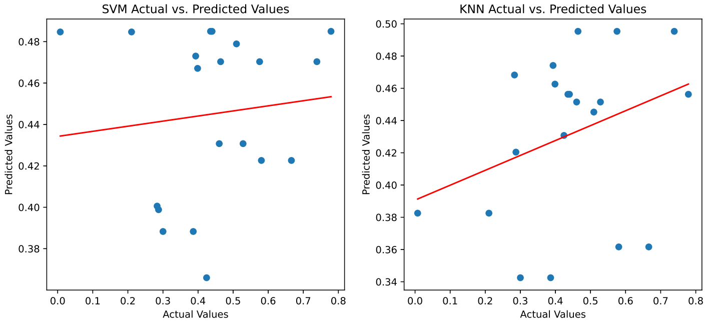

Figure 5: Correlation Plot on Actual and Predicted Values

Using a correlation heatmap to visualize the dataset, it was discovered that none of the input factors had a very strong correlation coefficient with the target variable. Specifically, the variables ”location_mention” and ”explicit” show a minor positive association, while the variable ”count” shows a slight negative relationship.

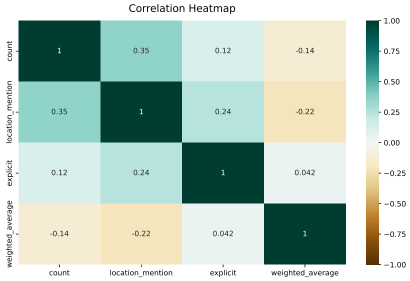

Figure 6: Correlation Heatmap

The Pearson correlation coefficients computed for the variables under examination provide insights into their relationships [location_mention : -0.215029], [explicit : 0.042363], and [count : -0.138480].

### 4.3 Training

Setting the optimal settings to train and test at an 80%-20% ratio, the classification results for all machine learning approaches are shown in the table below.

Table 1: SVM and KNN Classifier Performance

| | Accuracy | R-Squared | Mean Squared Error | Mean Absolute Error |
|---|---|---|---|---|
| SVM | 0.65 | 4.72 | 0.03 | 0.13 |
| KNN | 0.65 | 0.10 | 0.03 | 0.12 |

Table 2: Random Forest and XGBoost Performance

| | **Random Forest Performance** | | | **XGBoost Performance** | | |
|---|---|---|---|---|---|---|
| | Precision | Recall 3 | F1 | Precision | Recall 3 | F1 |
| Class 0 | 0.80 | 1.00 | 0.89 | 0.80 | 1.00 | 0.89 |
| Class 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Accuracy | | | 0.80 | | | 0.80 |
| Macro Avg. | 0.40 | 0.50 | 0.44 | 0.40 | 0.50 | 0.44 |
| Weighted Avg. | 0.64 | 0.80 | 0.71 | 0.64 | 0.80 | 0.71 |
| AUC Score | | | 0.66 | | | 0.66 |

Table 1 shows the performance SVM and KNN models, on the classification task. While SVM and KNN have equal accuracy (0.65), their performance differs in various ways. SVM outperforms KNN in terms of R-squared, with a score of 4.72 against 0.10. This suggests a considerably better fit to the data and greater explanatory ability for SVM. However, both models had the same mean absolute error (0.13) and mean squared error (0.03), indicating similar average prediction accuracy.

Table 2 compares the performance of the Random Forest and XGBoost model. Both the Random Forest and XGBoost models performed well on the classification task, with precision scores of 1.00 and recalls of 0.80 for class 0, indicating that 80% of class 0 samples were correctly identified without any misclassification of class 1 samples. Both models had an accuracy of 0.80, accurately categorizing 80% of the samples. Additionally, both models achieved a moderate AUC score of 0.66, demonstrating an adequate capacity to rank the classes. Despite these decent outcomes, there is still plenty of potential for improvement.

The assessment of songs in the predictive dataset constantly produces a value of 0, implying that their weighted averages are less than 0.5. This shows a flaw in the learning dataset and prediction model, as opposed to random guessing.

## 5 Limitations

The major limitation of this study is the variability of the dataset. The dataset consists of a limited number of known songs released in Nigeria in 2023. This small sample size may restrict the generalizability of our findings. The exclusive emphasis on a single year and country may trivialize temporal and geographical variations in music trends and audience preferences. Also, While efforts were made to ensure data accuracy, limitations in data collection methodologies may impact the result.

Table 3: Song Prediction Results

| Artist | **Track Name** | **Actual** | **SVM Predicted** | **KNN Predicted** |
|---|---|---|---|---|
| Chike | Ego Oyibo | 0.386 | 0.388 | 0.342 |
| CKay | Love Nwantiti | 0.528 | 0.431 | 0.451 |
| Olamide | New Religion | 0.399 | 0.467 | 0.463 |
| DJ Tarico | Yaba Buluku | 0.424 | 0.366 | 0.431 |
| Ruger | Bounce | 0.460 | 0.431 | 0.451 |
| Blaqbonez | Back in Uni | 0.393 | 0.473 | 0.474 |
| Rexxie | Abracadabra | 0.211 | 0.485 | 0.383 |
| Spyro | Only Fine Girl | 0.440 | 0.485 | 0.456 |
| Asake | Amapiano | 0.510 | 0.479 | 0.445 |
| Rema | Charm | 0.576 | 0.470 | 0.495 |
| Ayra Starr | Sability | 0.581 | 0.423 | 0.362 |
| Spyro | Who Is Your Guy? | 0.779 | 0.485 | 0.456 |
| Carterefe | Machala | 0.283 | 0.401 | 0.468 |
| Joeboy | Duffel Bag | 0.436 | 0.485 | 0.456 |
| Davido | Unavailable | 0.739 | 0.470 | 0.495 |

## 6 Conclusion and Future Works

This study looked into the relationship between musical success and the use of geographical mentions in song lyrics. A fresh dataset was created, which included references to physical locations and the frequency with which they were mentioned in each song in our corpus. To gauge a song's success, we designed a formula that combines Spotify's popularity % with TikTok's video count, using both existing popularity metrics from Spotify listeners and growing social trends.

We used visual analysis to display the distribution of geographical mentions and inferred the potentiality of musical success. Notably, our dataset includes common and collective place names such as ”club,” ”street,” ”party,” ”school,” ”outside,” etc., which lack accurate geo-referencing and hence were excluded from this study. Ultimately, based on the results of our analytical model, we conclude that, while the potential for music success is predictable, the sheer presence or frequency of geographical allusions is insufficient to forecast a song's success or social trend on its own. We regard this element as substantial, yet insufficient in independence.

The discovery that the relationship between geo-location mentions and music lyrics does not independently contribute to the prediction of music success is consistent with expectations, given the multifaceted nature of music production, which includes variables such as composition, acoustics, promotional strategies, marketing campaigns, and music videos.

Future work in this domain may need the use of a greater temporal scope in the analysis, allowing for the inclusion of a larger volume of data and a more diverse range of songs. Furthermore, future research may investigate the impact of sentiment on musical compositions, specifically the sentiments present in trenches or lush lyrical content. Additionally, efforts are being made to construct a corpus of low-resource African languages produced from musical compositions, providing researchers with accessible African datasets for innovative purpose.

## REFERENCES

Mark Altaweel. The geography of music. 1, 2020. URL https://www.geographyrealm.com/the-geography-of-music/.

AZLyrics. URL https://azlyrics.com.

Chelsea Bengier. The 50 best songs with a location in the title. *BestLife*, 2020. URL https://bestlifeonline.com/songs-with-a-location-in-the-title/.

Khaled Boughanmi and Asim Ansari. Dynamics of musical success: A machine learning approach for multimedia data fusion. *Journal of Marketing Research*, 58(6):1034–1057, 2021. doi: 10.1177/00222437211016495. URL https://doi.org/10.1177/00222437211016495.

O. Johansson and T.L. Bell. Sound, society and the geography of popular music. 2009. URL https://books.google.com.ng/books?id=ITicKKM8OMcC.

A. Leyshon, D. Matless, and G. Revill. The place of music. 1998. URL https://books.google.com.ng/books?id=q2Jt4r4pnf4C.

Genius Lyrics. a. URL https://genius.com.

Song Lyrics. b. URL https://songlyrics.com.

Sean H Merritt, Kevin Gaffuri, and Paul J Zak. Accurately predicting hit songs using neurophysiology and machine learning. *Front. Artif. Intell.*, 6:1154663, June 2023.

Dmitry Pastukhov. What does the future hold? 4 music trends expert predictions for 2023 (and beyond). 2022. URL https://soundcharts.com/blog/music-industry-trends.

T. Pinch and K. Bijsterveld. The oxford handbook of sound studies. 2012. URL https://books.google.com.ng/books?id=KuRfLG0IedYC.

George Revill. Geographies of music, sound, and auditory culture. *Oxford Bibliographies*, 2017. URL https://www.oxfordbibliographies.com/view/document/obo-9780199874002/obo-9780199874002-0130.xml.

SongStats. URL https://songstats.com.

Spotify. URL https://spotify.com.

J. Sterne. The sound studies reader. 2012. URL https://books.google.com.ng/books?id=kbufJrm6c9YC.

TikTok. URL https://tiktok.com.

Michael E.W. Varnum, Jaimie Arona Krems, Colin Morris, Alexandra Wormley, and Igor Grossmann. Why are song lyrics becoming simpler? a time series analysis of lyrical complexity in six decades of american popular music. *PloS one*, 16(1), January 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0244576. Publisher Copyright: © 2021 Varnum et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Ryan Werth. Music matching with natural language processing. 2021. URL https://medium.com/analytics-vidhya/music-matching-with-natural-language-processing-44d5e50c9b9f.