Linguistic Identity in Music: Classification of Song Lyrics Written by Native vs. Non-Native English Speakers

Adela-Nicoleta Corbeanu
adela_uchiha@yahoo.com

Abstract

This study explores the classification of English-language song lyrics based on whether they were written by native or non-native English speakers. Given the artistic and expressive nature of song lyrics, this problem presents unique linguistic and cultural challenges that differ from traditional Native Language Identification (NLI) tasks. Using a carefully-curated dataset of $\sim$5000 songs, various machine learning techniques were employed, including supervised and unsupervised learning approaches. The best performance was achieved by a transformer-based approach, which obtained an accuracy of 0.817. Key linguistic differences were identified, such as the prevalence of colloquial expressions in native lyrics. The findings suggest that while linguistic features provide valuable cues, cultural and stylistic factors also play a crucial role in distinguishing native from non-native songwriting styles.

1 Introduction

The problem I am trying to solve is classifying English-language songs lyrics based on whether they were written by native or non-native English speakers.

I chose this problem because it offers an interesting intersection of language, culture, and music. Song lyrics are often deeply tied to the songwriter's cultural and linguistic background, and differentiating between native and non-native language usage in song lyrics provides an opportunity to explore how language proficiency influences artistic expression. Additionally, I am very interested in understanding language patterns in artistic writing. Songs lyrics serve as a rich, varied source of data for this kind of analysis, and are also fairly easy to collect.

2 Related Work

While I am not classifying the native language itself, I believe the task still falls under the umbrella of Native Language Identification (NLI), a common NLP task that tries to predict the native language of a person based on a text written by them in a second language (most commonly English).

Probably the most popular dataset for NLI is the ICLE (The International Corpus of Learner English) (Granger et al., 2009), a proprietary (not freely available) dataset consisting of English essays written by advanced students from many different backgrounds.

Very similar to ICLE, TOEFL11 (Blanchard et al., 2013) is a dataset consisting of $11,000$ English essays written by non-native speakers from 11 different backgrounds. This dataset has been used as support in an NLI shared task in 2013, where the most successful approach obtained a score of 83.6% accuracy (Jarvis et al., 2013). Despite this significant achievement, the winning team concluded that the language alone does not solely determine the author's origin, but that culture plays a crucial role as well.

The difference between these aforementioned datasets and my dataset lies in the type of texts: academic essays written by students vs. song lyrics. While the essays might contain mistakes and might follow the same structure, song lyrics are written with an artistic purpose and liberty of structure and are carefully composed during a long period of time, as opposed to spontaneous essays during exams.

3 Methodology

3.1 Dataset

The dataset is called *English-Lyrical-Origins* (Corbeanu, 2024) and it consists of the lyrics of $4,941$ different songs, belonging to 86 artists (seen in 3, 2) of 34 nationalities (seen in 1).

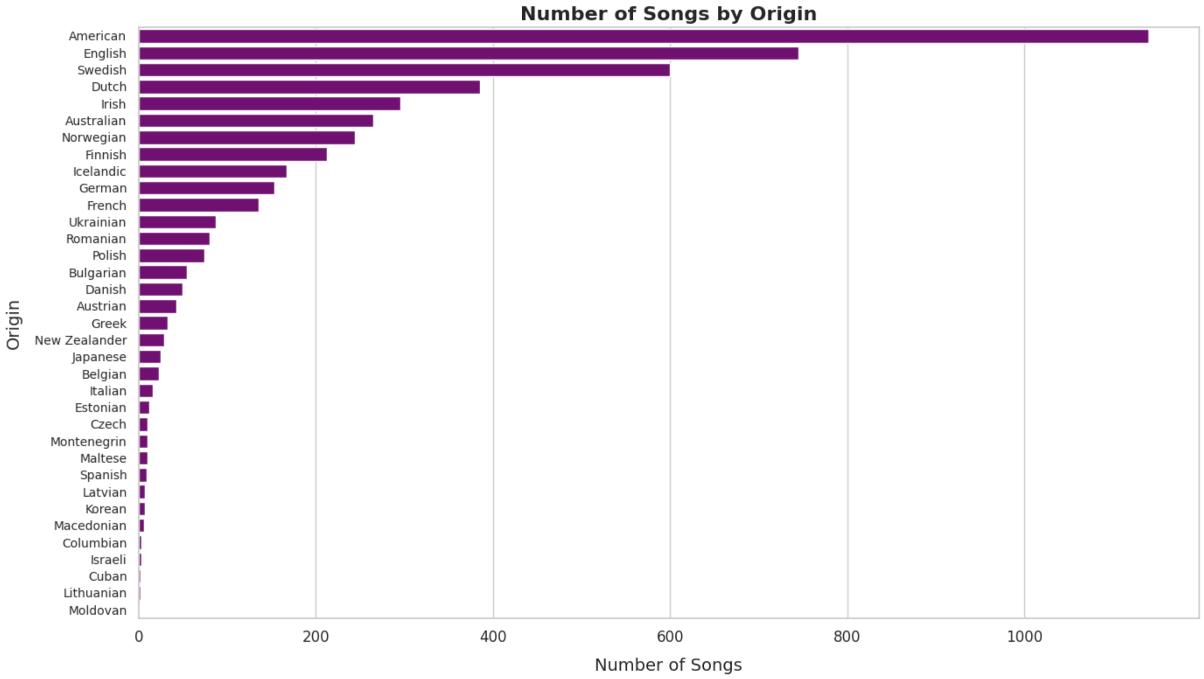

Figure 1: Number of songs of each nationality in the dataset.

Each song in the dataset has the following fields:

- `title`: The title of the song.
- `artist`: The name of the performing artist or band.
- `lyrics`: The full lyrics of the song.
- `native`: Indicates whether the artist is a native English speaker. (can be "yes" or "no").
- `origin`: The nationality of the artist.

The lyrics were obtained via web scraping and were annotated based on the artist's nationality. The artists were manually investigated and selected based on these criteria:

- They widely claim they write their own songs, at most receiving help from same-nationality people.
- They have released songs in English.
- (for non-natives) They were born and raised in a non-English-speaking country, in a non-English-speaking household.
- (for natives) They were born and raised in an English-speaking country and household.

The lyrics on the website are not provided by the official artists, but are instead transcriptions made by individuals who have listened to the song.

When collecting the data, I aimed for maximum diversity, ensuring an equal balance of 50% native and 50% non-native English songs, including a variety of nationalities among the native songs. For non-native songs, I aimed to include data from all continents. However, due to the influence of the Eurovision Song Contest, which served as a valuable source for identifying artists, a big part of the songs ended up being from Europe.

The data is shuffled and it has two predefined splits: *train* ($3,941$ songs) and *test* ($1,000$ songs).

By looking at the lyrics, I would not expect the classification to be trivial. The one difference that seems to stand out to me is that English-native songs use more slang and idioms, such as: "taking me places", "ya" instead of "you", "lovey dovey", etc.

One point to keep in mind is that all these artists, both native and non-native, come from all kinds of educational backgrounds, so the task should not be looked at in the manner of identifying who speaks better English, for example: Iggy Azalea (English native from Australia) dropped out of school early, while multiple non-native European singers pursued advanced formal education. I believe it is rather the case of distinguishing formally-learned

English (the case of non-native artists) from naturally picked-up English (the case of native speakers).

Although it would be of interest to compute readability tests such as Flesch-Kincaid, they all include the number of sentences in their formula. This is problematic in the case of lyrics, where the text is not necessarily separated into sentences and there is no trivial workaround. Therefore, we must limit ourselves to:

$$\text{TypeToken\_Ratio} = \frac{\text{Unique\_words}}{\text{Total\_words}}$$

The average TTR for English-native songs is 0.3502, while the average for non-English-native songs is, surprisingly, 0.4102.

When looking at the lexicon count (how manu unique words) each song has, we obtain the following statistics:

Metric	Native	Non-Native
Max Lexicon Count	3108	885
Avg Lexicon Count	341	196

3.2 Preprocessing

The dataset specifies a mandatory preprocessing step when used for classification tasks: deleting all text inside brackets, as it may contain data that gives away the label, example: `[Chorus - John Lennon]`, `[Verse 1 - Eminem]`. This needs to be done to ensure fairness and relevance of results.

Secondly, I found casing not to be relevant in this task. This makes sense if we think that the casing is done by the individuals who transcribed the songs, not by the artists themselves, so the casing style has nothing to do with the artists.

Other preprocessing techniques I tried are removing punctuation and stopwords. Punctuation had very insignificant impact (just like casing, it is not written by the artists), while the removal of stopwords actually impacted the score *negatively*, resulting in $\sim 3\%$ smaller accuracy.

Stemming and lemmatization lowered the results even more, negatively impacting with $\sim 5\%$ accuracy.

3.3 Method

3.3.1 Baseline

The baseline has been computed using the very simple probabilistic classifier Naive Bayes, obtaining an unexpectedly high accuracy of 0.702.

3.4 Principal Component Analysis

I have used PCA to analyze the most representative features.

Lyrics most aligned with the first principal component:

- *lost nowhere searching home still turning past gone time omen showed took away preparations done can...*
- *final destiny sunrise never came still night lamp never faded away farewell word afterglow brave mor...*
- *cleared fog veiled around blurred sights suddenly im longer aching honor plights rising moon skin pe...*
- *beloved name inside heart fleeting glance became start missing word still awaiting wretched deceptio...*
- *abode mongst stars waiting long enough last breath life stare nothing right times resembling devils...*

Lyrics least aligned with the first principal component:

- *mo bounce bounce bounce bounce bounce bounce bounce bounce bounce motherfuckin house mo bounce mothe...*
- *get want like click want pic like click cheers glass like click cash register goes click cant fuck c...*
- *intro tony yayo 50 cent shady yeah run know actin like dont know run yeah know actin like dont know...*
- *uhoh runnin breath oh got stamina uhoh running close eyes well oh got stamina uhoh see another mount...*
- *told boy kiss girl take trip around world hey hey bop shuop mbop bop shuop hey hey bop shuop mbop bo...*

By intuitively analyzing data in relation to the first principal component, it seems that it avoids “slangy” and colloquial terms, which are normally not present in literature or other common sources for learning English, pointing therefore towards native speakers. Some examples are the use of interjections such as “yeah”, as well as the auxiliary verb “got”.

Common words in first-component lyrics: *still, would, past, away, cant, last, around, never, word, rising*

Common words in non-first-component lyrics: *bounce, dont, yeah, like, know, give, greatest, im, got, boys*

3.4.1 Supervised learning

I have tried multiple supervised learning approaches. While the transformer-based approach score the best, it was not an absolute victory, but rather a relatively small difference.

Multilingual BERT

The Multilingual BERT model achieved the best performance with an accuracy of 0.817, outperforming other models in the evaluation. Furthermore, the model demonstrated an average confidence of 0.9341, reflecting its strong predictive certainty across its predictions. The combination of high accuracy and confidence suggests that Multilingual BERT not only performs well but also provides reliable and consistent results, making it a robust choice for this task.

If training on song titles instead of lyrics, the model obtains 0.635 accuracy, which goes down abruptly to 0.5 (same as random choice) if stopwords are removed. We may conclude that titles are not so informative.

Support Vector Machine

As part of the aforementioned 2013 NLI shared task, 7 out of the top 10 teams used variations of SVM’s (Zampieri et al., 2017). Particularly, most of them used word n-grams and character n-grams. I experimented with their n-grams values, but the results were lower when used on lyrics than on essays. However, using character 7-grams does in fact obtain the highest accuracy, 0.796.

Random Forest Classifier I have noticed that the Random Forest model obtains between 73-75% accuracy without much effort. When using character 7-grams, TF-IDF vectorization, and fine-tuned hyperparameters, it goes to 0.779 accuracy.

Some of the most relevant n-grams for the model turned out to be: *ain’t*, *’cause*, *’em*. I would say this is predictable, because they belong to natural, informal speech, which is not common for people who learned English formally.

3.4.2 Unsupervised learning

While unsupervised learning is not common for these kind of tasks, I decided to give it a shot nonetheless.

K-Medoids

K-Medoids is a clustering-based unsupervised learning algorithm that, combined with PCA and Sentence-BERT embeddings, obtained an accuracy of 0.637:

Metric	Non-Native	Native
Precision	0.645	0.629
Recall	0.608	0.666
F1 Score	0.625	0.647
Accuracy	0.637	0.637

Self-Organizing Map

SOM is a grid-based unsupervised neural network that organizes data based on similarity. Unlike traditional neural networks that rely on error correction, SOMs use competitive learning, where neurons compete to represent input patterns. The training process adjusts neuron weights iteratively, ensuring that similar data points are mapped closer together. I used SOM for separating native and non-native songs into clusters, and it performed similarly to K-Medoids:

Metric	Non-Native	Native
Precision	0.591	0.743
Recall	0.860	0.404
F1 Score	0.700	0.523
Accuracy	0.632	0.632

4 Future Work

Future work should aim to address the limitations identified in this study. I also believe that with enough time and analysis, the classification performance can be improved.

5 Conclusion

This research demonstrates that machine learning models can differentiate between native and non-native English song lyrics with notable accuracy. The results indicate that native English speakers tend to use more informal, idiomatic, and colloquial expressions. The findings contribute to the broader field of NLI by highlighting how artistic expression interacts with formal language-learning. While the Multilingual BERT model performed best, other models, such as SVM and Random Forest, also provided competitive results.

Limitations

While the dataset was created with diversity in mind, some nationalities and artists might still be over/under-represented. Over-representation might happen particularly English-native nationalities, given that there is less of them compared to non-native ones. This might also happen to some artists.

Secondly, there is no possible way of accounting for the phenomenon of ghostwriting, which would probably impact non-native songwriters more (they could use native writers, while the opposite is unlikely). However, I would expect it to be more common among bigger artists, which is rarely the case for the non-native artists in this dataset.

Furthermore, while some linguistic features were analyzed, the study did not account for deeper stylistic elements, such as rhyme schemes, metaphors, or genre-specific phrasing, which could influence classification.

Ethical Statement

One possible unethical use of this research could be leveraging it to stereotype or discriminate against non-native English-speaking artists. To ensure ethical usage, researchers utilizing this dataset should be mindful of its limitations and avoid making generalized conclusions about language proficiency or artistic value.

Personally, I believe that while machine learning can uncover linguistic trends, it is crucial to acknowledge the broader cultural and artistic contexts when interpreting results.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013:i–15.

Adela Corbeanu. 2024. English-lyrical-origins dataset. https://huggingface.co/datasets/AdelaCorbeanu/English-Lyrical-Origins. Accessed: 2025-02-02.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English (Version 2)*. Presses universitaires de Louvain.

Scott Jarvis, Yves Bestgen, and Steve Pepper. 2013. Maximizing classification accuracy in native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–118, Atlanta, Georgia. Association for Computational Linguistics.

Marcos Zampieri, Alina Cristea, and Liviu Dinu. 2017. Native language identification on text and speech. pages 398–404.

A Additional dataset diagrams

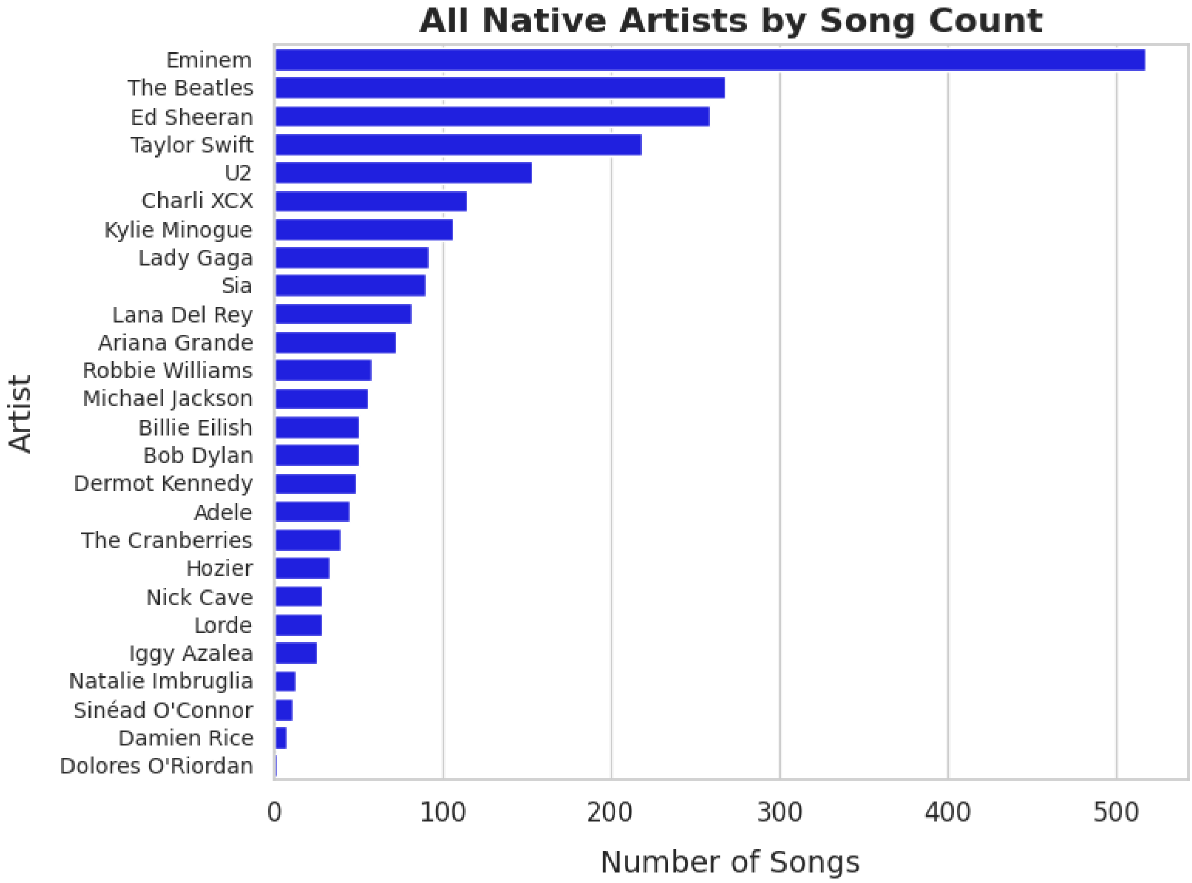

Figure 2: Number of songs per artist for English native artists.

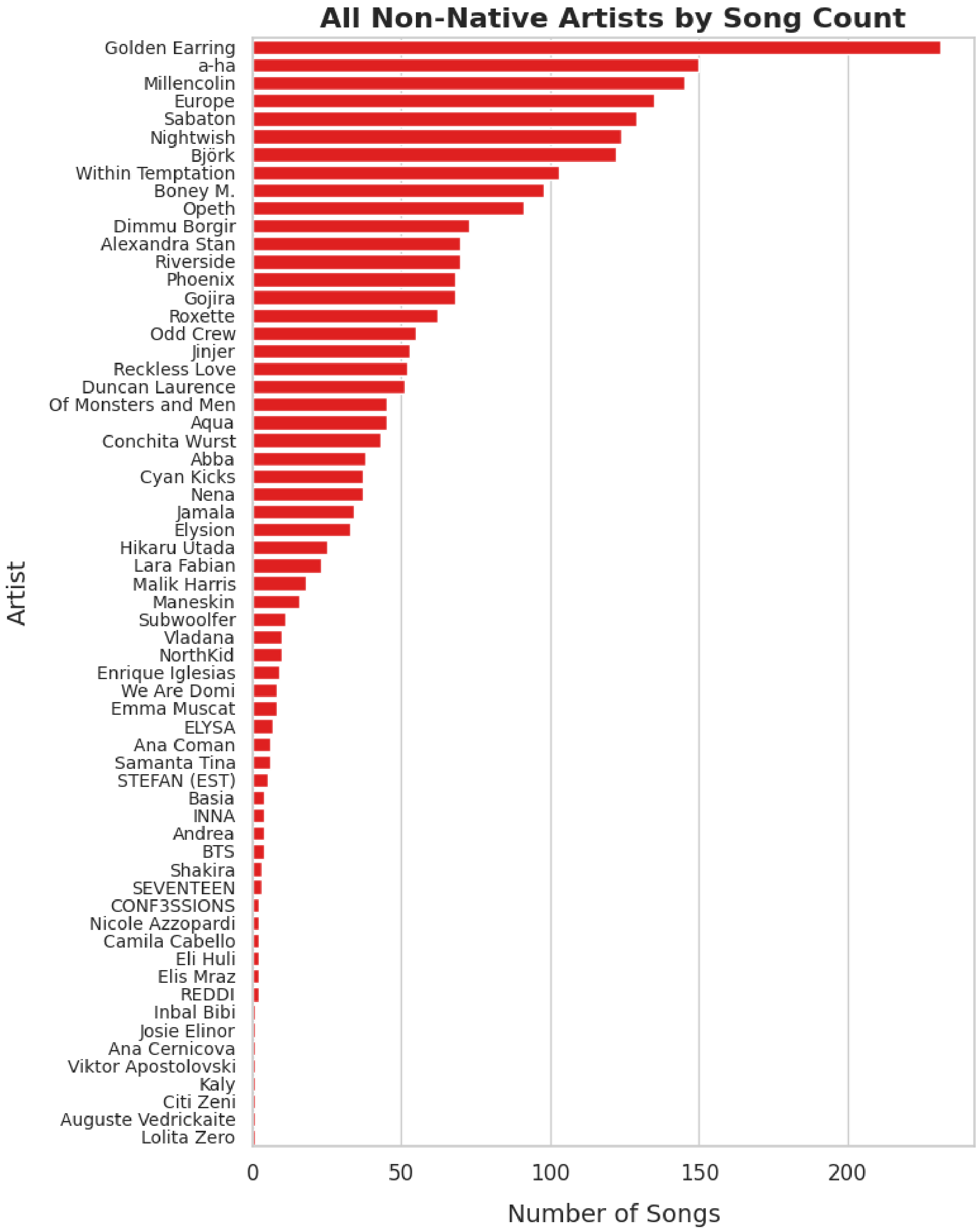

Figure 3: Number of songs per artist for non-native artists.