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

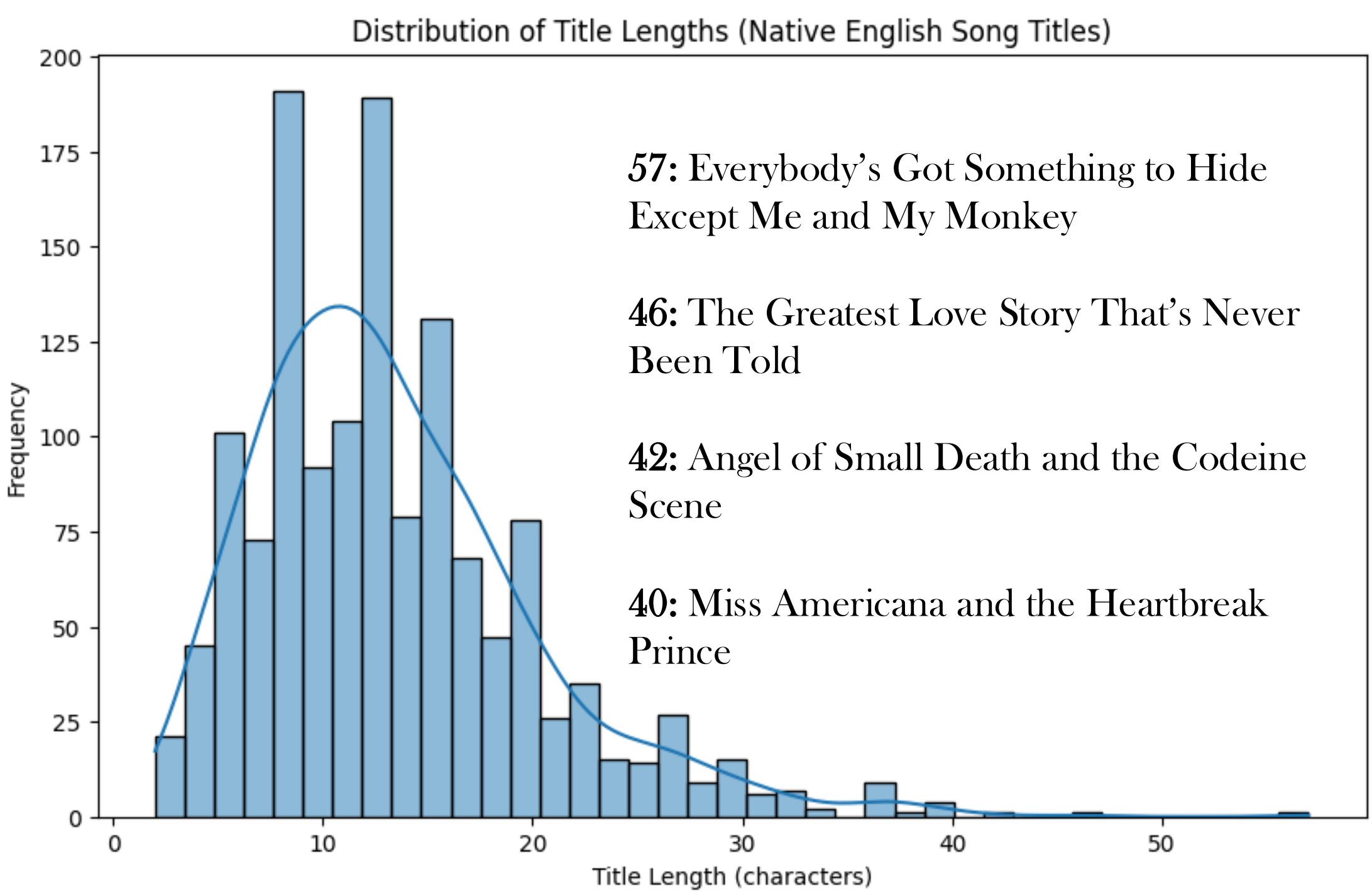
Adela-Nicoleta Corbeanu, 412

February 2025

♫ *English-Lyrical-Origins* Dataset ♫

https://huggingface.co/datasets/AdelaCorbeanu/English-Lyrical-Origins

title string · *lengths* 32↔38 1.2%	**artist** string · *classes* 58 values	**lyrics** string · *lengths* 0 52.6k	**native** string · *classes* 2 values	**origin** string · *classes* 32 values
Public Service Announcement 2000	Eminem	[Announcer: Jeff Bass & Eminem] This is another public service announcement…	yes	American
Merry Christmas Mr. Lawrence - FYI	Hikaru Utada	[Intro] Huh, huh, huh, huh Huh, huh, huh-uh-uh Huh, huh, huh, huh Huh, huh, huh-uh…	no	Japanese
Give That Wolf A Romantic Banana	Subwoolfer	[Intro] Oh-ooh-ooh-ooh-ooh-ooh [Verse 1] Not sure I told you, but I really like…	no	Norwegian
Having Grandma Here for Christmas	Subwoolfer	[Intro] Deck the moon with lots of oldies Yum-yum-yum-yum-yum, yum-yum, yum, yum…	no	Norwegian
Do You Know? (The Ping Pong Song)	Enrique Iglesias	[Intro] Do you know? Do you know? Do you know? [Chorus] Do you know what it feels…	no	Spanish
Mary Ellen Makes the Moment Count	a-ha	Mary cries out, "For the love of God" As she's walking out the laundromat Down the…	no	Norwegian

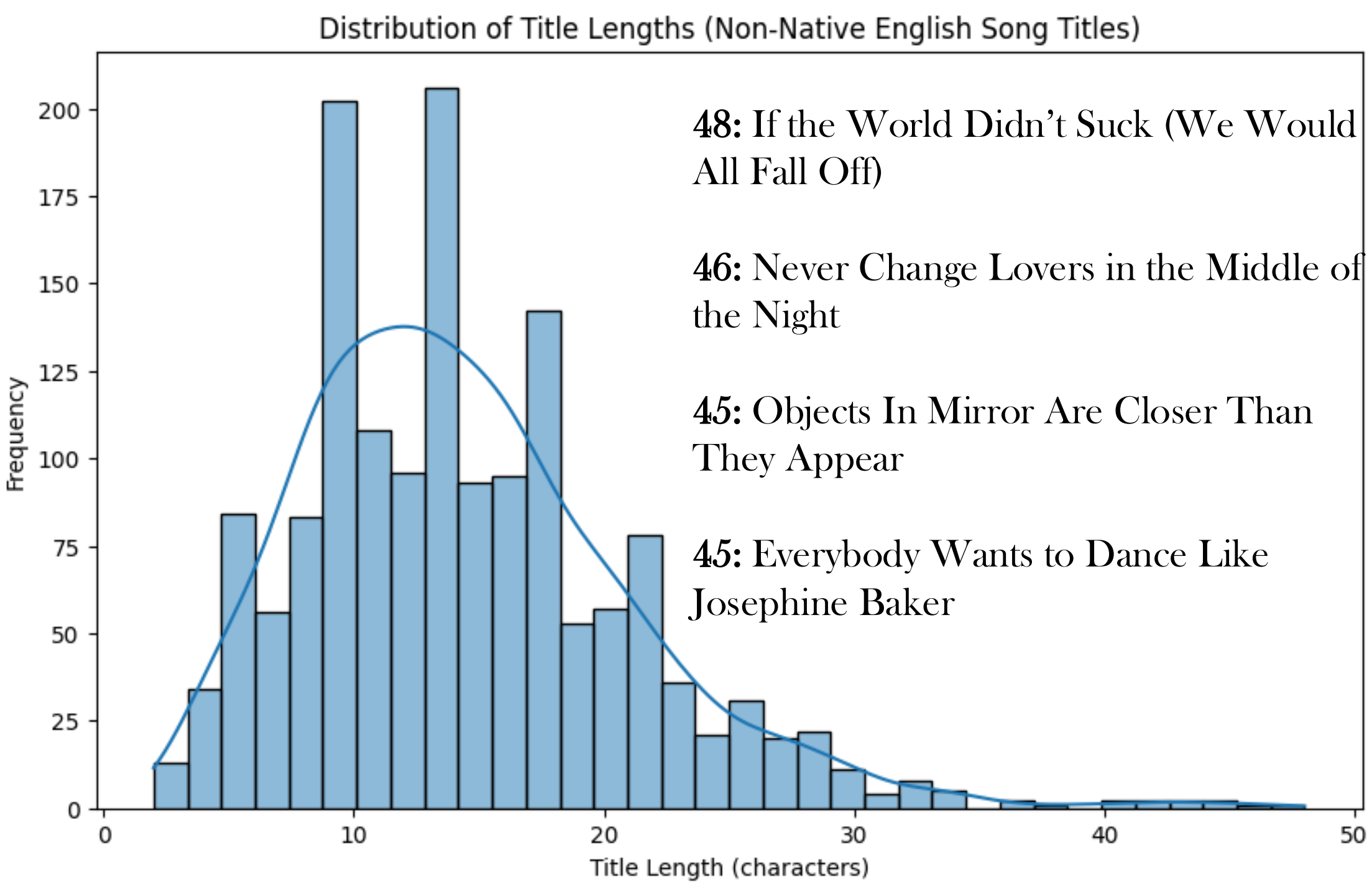


57: Everybody's Got Something to Hide Except Me and My Monkey

46: The Greatest Love Story That's Never Been Told

42: Angel of Small Death and the Codeine Scene

40: Miss Americana and the Heartbreak Prince



48: If the World Didn't Suck (We Would All Fall Off)

46: Never Change Lovers in the Middle of the Night

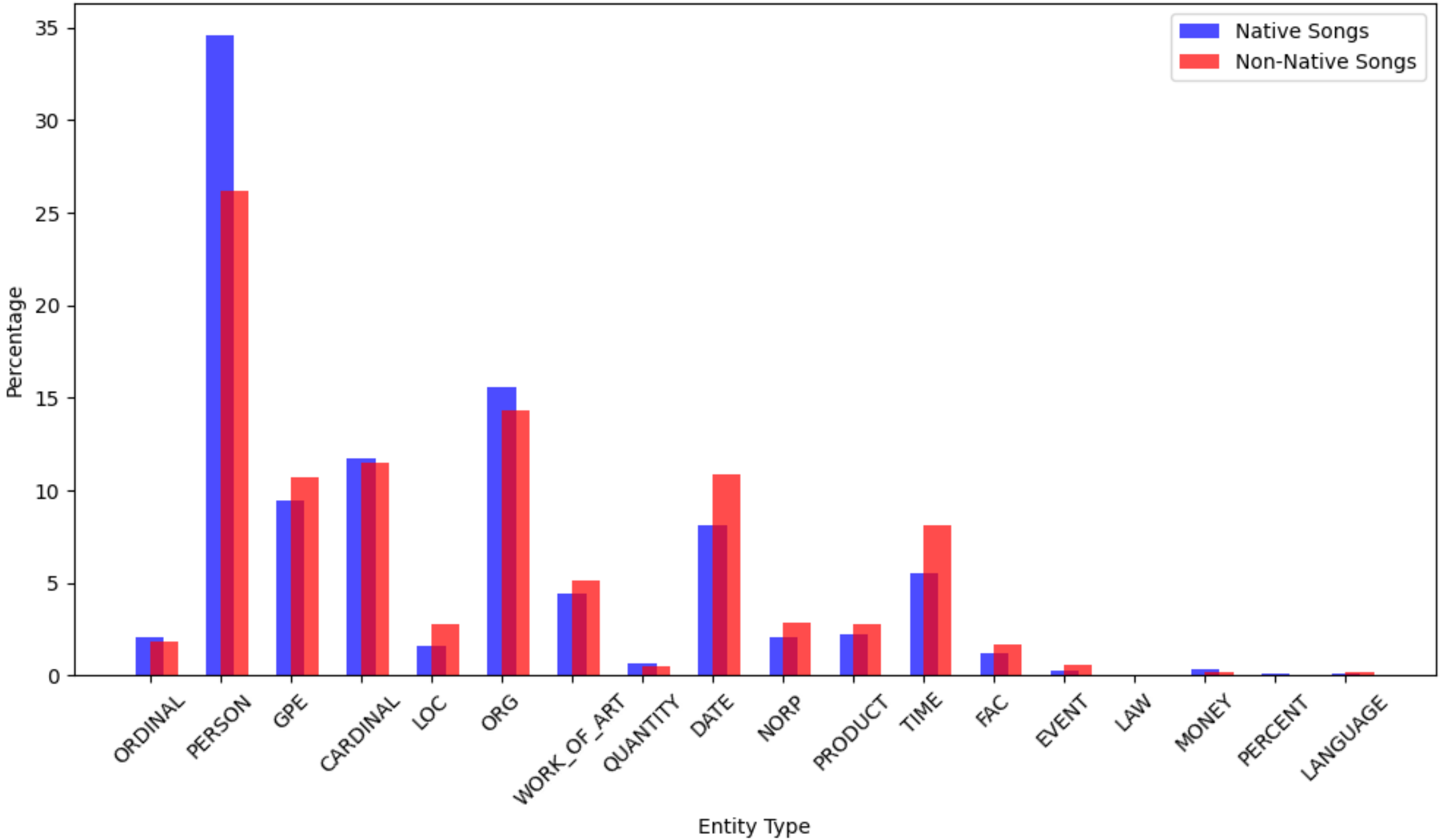
45: Objects In Mirror Are Closer Than They Appear

45: Everybody Wants to Dance Like Josephine Baker

Lexicon counts

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

For comparison:
"Luceafărul" by Mihai Eminescu
(English version) has
`lexicon_count=2201`


Named Entity Distribution in Native and Non-Native Songs
Native Songs
Non-Native Songs
Percentage
0
5
10
15
20
25
30
35
ORDINAL
PERSON
GPE
CARDINAL
LOC
ORG
WORK_OF_ART
QUANTITY
DATE
NORP
PRODUCT
TIME
FAC
EVENT
LAW
MONEY
PERCENT
LANGUAGE
Entity Type

The 2013 Native Language Identification shared task

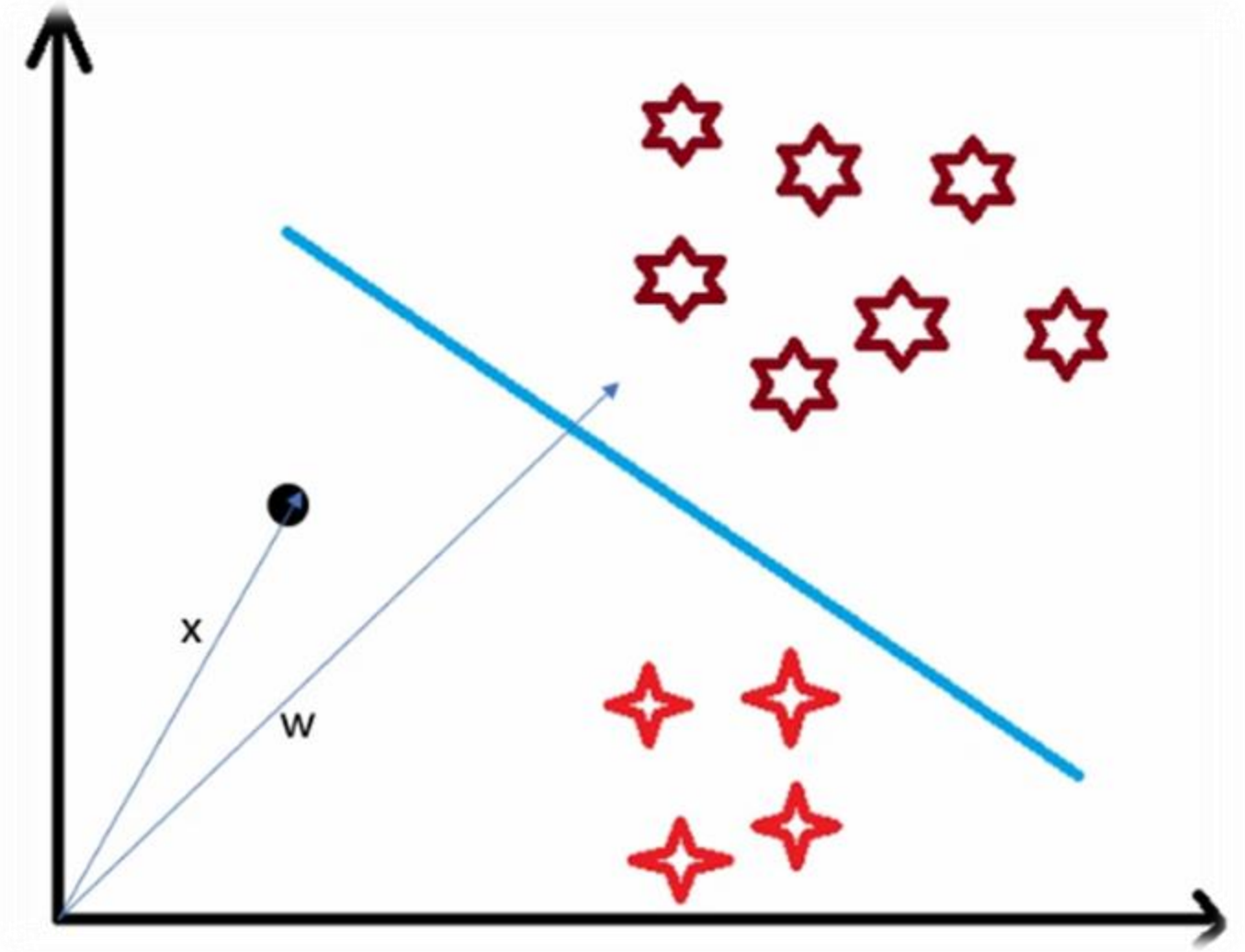
Multilingual BERT

- 81.7% accuracy
- Average confidence: 0.9341

Lowest Confidence Prediction:

- **Text:** got heels high satellite got tip toe walking moonlight whos whos whos busy busy cares dont youre better walk hit lights fingerlicking good tricking bouncers good whos love whos whos itty bitty whos drunk
- **Predicted Label:** yes
- **True Label:** no
- **Confidence:** 0.5011

Support Vector Machine



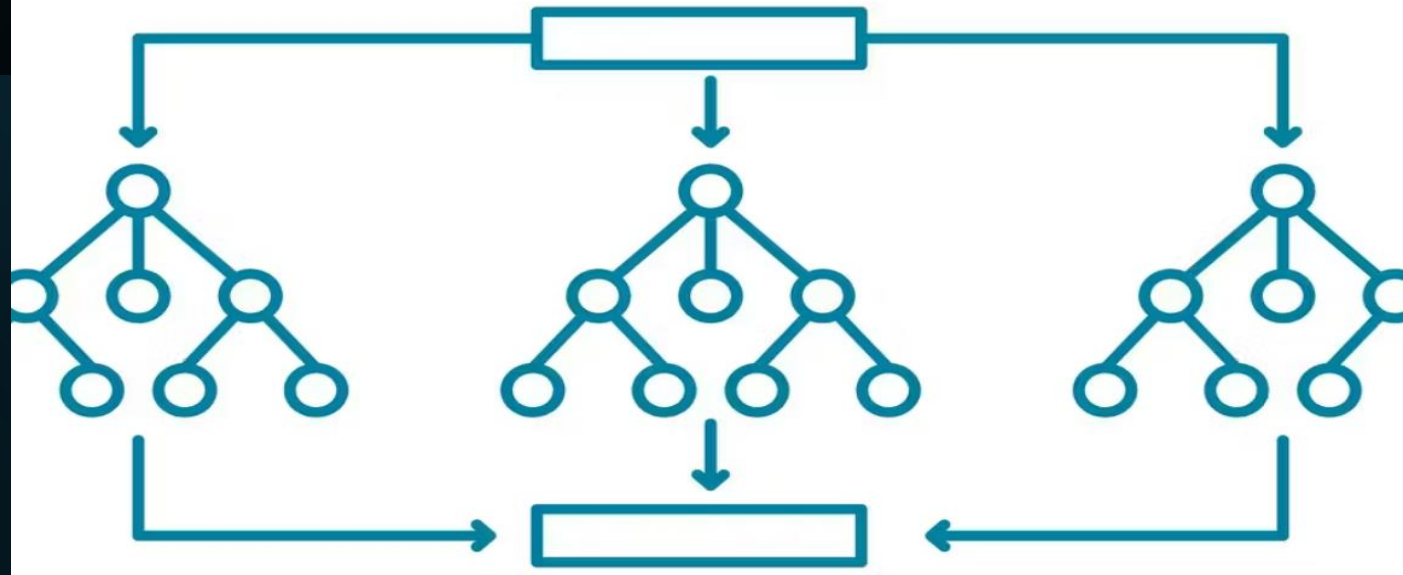
- 79.6% accuracy
- The most promising tried model

Random Forest classifier

Accuracy on lyrics: 77.9%

Accuracy on titles: ~55%

- ain’t
- ’cause
- ’em
- char 7-grams
- TF-IDF




Thank you!