THE INFLUENCE OF ENGLISH LANGUAGE WORDS IN TAMIL FILM SONG LYRICS CORPUS: A MACHINE LEARNING APPROACH

Sabananthan Vijayarangan[1], Freya Russell[2], Tilak Ginige[3], Iain Green[4]
[1] Independent Researcher, India
[2, 3, 4] Department of Life & Environmental Sciences, Faculty of Science & Technology, Bournemouth University, Talbot Campus, Fern Barrow, BH12 5BB

[1] v.sabananthan@gmail.com, [2]frussell@bournemouth.ac.uk, [3]tginige@bournemouth.ac.uk, [4]igreen@bournemouth.ac.uk

Corresponding author email: v.sabananthan@gmail.com

***Abstract*—** The use of English words in Tamil lyrics has significantly increased in recent years due to globalisation, Western cultural exposure, and the evolution of the Tamil music industry. Traditionally, Tamil lyrics used pure Tamil words, which were well-received in the early 1960s. However, the younger generation's exposure to Western influences and English-speaking media has led to a greater incorporation of English words in Tamil lyrics, reflecting broader cultural and societal changes. This blending of languages has made it difficult to classify the language of these compositions accurately. To address this, this study proposes a machine learning model to handle code-switching and multilingual text for identifying English words in Tamil lyrics. The objectives include testing the hypothesis that modern lyrics contain more English words than earlier compositions and using an unsupervised machine learning algorithm to cluster lyricists based on their use of English and other foreign language elements. By leveraging natural language processing (NLP) methods and machine learning techniques, the study aims to analyse language usage patterns in Tamil lyrics and explore the relationship between language, culture, and society in contemporary music.

***Keywords* — Machine learning; Tamil; Music; Globalisation; Algorithm; Culture; Code-Mixing**

1. Introduction

Six decades ago, Tamil film song lyrics were crafted exclusively using pure Tamil words by Tamil lyricists. Due to globalisation, Indians began pursuing education and career opportunities abroad, leading to exposure to various cultures and the assimilation of English as a global language. Consequently, Tamil lyrics have evolved, incorporating changes in the language, Western cultural influences, and societal developments.

The evolution of Tamil film songs reflects the globalised world's changing dynamics. A comprehensive dataset has been compiled, consisting of words from Tamil, English, code-mixed English-Tamil expressions, and other languages. This multilingual mix challenges traditional language identification methods. Advanced language models capable of handling code-switching, such as code-mixed or multilingual models, are more effective in identifying and interpreting the diverse language patterns in Tamil songs. Code-mixing, the practice of combining two or more languages in communication, often involves using words or phrases from other languages alongside Tamil. This is prevalent among Tamil speakers, especially in urban areas [1]. While natural, code-mixing can pose literacy challenges, such as ambiguity and comprehension difficulties for readers unfamiliar with the mixed languages. It can also impact the development of standard Tamil language skills, particularly in children exposed to code-mixed languages early on [2]. However, code-mixing can

address lexical gaps in Tamil and enhance its richness, serving as a tool for language learning. songs.

This study focuses on extracting features like English words, Tamil code-mix suffixes, and Tamil words, and identifying lyricists who use foreign language words in Tamil lyrics. It employs data science techniques such as POS tagging, the bag of words method, TF-IDF, and synset methods to identify foreign words. The CRF algorithm, used for sequence tagging tasks, will be explored for language identification. Additionally, the K-Means algorithm will be used to group lyricists based on their use of foreign words, offering insights into their linguistic tendencies. The preference for K-Means is due to time constraints and the lack of data for supervised training.
The study does not include identifying languages other than Tamil and English, or meaningless words within the dataset. Semantic validation of transliterated words and exploring algorithms other than CRF and K-Means are also excluded due to runtime and resource constraints.

This research aims to explore code-mixing in Tamil song lyrics and its impact on the use of foreign words, specifically English. By applying data science techniques and clustering algorithms, it seeks to identify lyricists who incorporate foreign words into their Tamil lyrics, contributing to the academic understanding of code-mixing and offering practical applications in natural language processing, cultural studies, and multilingual communication.

2. Materials and Methods

A dual approach combining classification and clustering methods is utilised to analyse the Tamil lyrics corpus. The Conditional Random Fields (CRF) algorithm is used for its sequence modelling capabilities, essential for text classification, employing features like Part-of-Speech (POS) tags and Inside-Outside-Beginning (IOB) labels to precisely identify language entities. The K-Means clustering algorithm is chosen to explore patterns in foreign language words and Tamil code-mix suffixes among lyricists, grouping them based on similarities in their word usage patterns.

The process shown in Figure 1 involves several steps to ensure model accuracy and robustness. It starts with data preparation and cleaning, where the raw Tamil lyrics are processed to remove any extraneous characters and inconsistencies. The cleaned text is then tokenised into individual words. Feature engineering follows, extracting relevant attributes

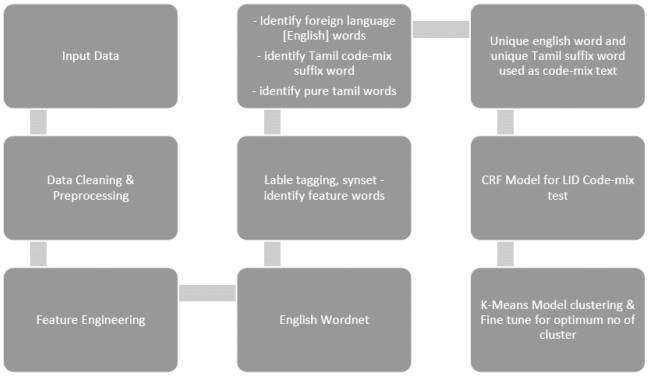

Figure 1 Flow diagram of the process used.

like lexical features and POS tags for both CRF and K-Means algorithms. IOB labeling identifies language entities such as Tamil and English words and code-mix suffixes, providing the ground truth for training the CRF model. Unique words are identified and counted, serving as features for both algorithms. The CRF model is then applied to identify language entities using the IOB labels and feature-engineered data, with its performance assessed through precision, recall, and F1-score metrics. Finally, the K-Means algorithm clusters lyricists based on their code-mix word usage, using the unique word counts and other features to group them and reveal patterns or trends in their use of code-mixed language.

2.1. Classification

The use of Part-of-Speech (POS) tagging and Inside-Outside-Beginning (IOB) labeling lays a robust foundation for applying the Conditional Random Fields (CRF) model [3] to Tamil lyrics. The process begins with data cleaning and preprocessing, which involves breaking the text into tokens and removing

unwanted characters. POS tagging assigns grammatical roles to words, and IOB labeling identifies language entities, crucial for understanding code-mixing. Feature engineering enhances the dataset by including attributes such as POS tags, IOB labels, and word patterns, essential for training the CRF model. The CRF model is then trained on these features to detect and analyse code-mixed words, with its performance evaluated through precision, recall, and F1-score metrics.

2.2. **Clustering**

For the first time, this paper employs the K-means algorithm [4] to cluster lyricists based on their usage of code-mix words in Tamil lyrics. To determine the optimal number of clusters, both the elbow curve and silhouette score methods are used. The silhouette score assesses how well data points fit within their assigned clusters compared to other clusters, with higher values indicating better clustering. The elbow curve evaluates the sum of squared distances within clusters to help identify the optimal number of clusters. The machine learning models are trained on a labelled dataset where each word in the Tamil lyrics is tagged as either 'Tamil' or 'English'. This allows the model to learn patterns for predicting the language of new, unseen data. Human experts annotate the data, providing the correct labels for training. After training, the model's predictions are evaluated and compared with actual outputs to assess its accuracy and make necessary improvements.

2.3. **Resource Requirements**

Creating an effective research environment is crucial for achieving successful results. This involves both tangible elements, like infrastructure and computing power, and intangible aspects, such as selecting appropriate tools and Python packages. Adequate infrastructure enhances research efficiency by supporting data processing and analysis, leading to meaningful insights. For this research, ensuring the computing environment and tools are suitable is essential for effective data handling and analysis.

2.2.1 **Hardware**

Typically, High processing power computer high memory, and storage capacity can handle data processing and modelling tasks effectively. Table 1 below provides configuration list which makes the processing and modeling tasks manageable while also considering cost and resource optimization if used in cloud.

Table 1 Hardware Resources Required

Hardware	
Machine	Any compute with xlarge or higher in size preferably cloud/on premise
OS	Windows/Linux
Processor	16-32 core CPU
RAM	64GB
System Type	64-Bit
Network Bandwidth	10 Gbps
IOPS	16,000 IOPS
Storage	50 -100 GB HDD/SDD
Graphics device	VGA (1024 x 768)
Ethernet adapter	Min 1 gigabit per second

2.2.2 **Software**

Software packages like Jupyter Notebook is a widely used tool that provides a user-friendly and interactive environment for writing and executing Python code. Table 2 presents lists of Python packages, which serve diverse purposes, such as storing, retrieving, querying the input dataset, build model and evaluate its performance. Version control systems, such as Git enable tracking any modifications in the code, thus providing an overview of what changes were made and by whom.

Table 2 Software Resources Required

Software	
Operating System	Windows (or) Linux,
Python	V3.7 / V3.8
PIP	V23.1.2
Google Colab notebook	Latest version

Jupyter notebook	Latest version
pandas	Latest version
numpy	Latest version
matplotlib	Latest version
seaborn	Latest version
Sklearn	Latest version
Statsmodels	Latest version
scipy	Latest version
re	Latest version
plotly	Latest version
nltk	Latest version
wordcloud	Latest version
59klearn_crfsuite	0.3.6
pyenchant	Latest version
Linux package: enchant-2	Latest version
English wordnet(Synset)	Latest version
Version Control Systems (Git)	Latest version

3. Results and Discussion

The interweaving of English words with Tamil creates a vibrant and complex linguistic pattern, reflecting cultural nuances and artistic expressions. This research systematically unravels these patterns through Exploratory Data Analysis (EDA) and the application of a Conditional Random Field (CRF) and K-Means clustering model. The initial phase involved an in-depth EDA, providing a foundation for understanding the dataset's structure and characteristics. Key preprocessing steps, such as tokenization and special character removal, revealed essential insights. The discovery of trends, patterns, and anomalies offered a preliminary understanding of the complex interplay of languages within the Tamil lyrics corpus.

Building on the insights from EDA, a CRF model was meticulously designed, employing techniques like Part-of-Speech (POS) tagging and Inside-Outside-Beginning (IOB) labeling. The CRF model was trained and evaluated using metrics such as precision, recall, and F1-score to gauge its performance in identifying code-mixed words. The Silhouette Score provided insights into how well the K-means clusters were defined. This research aligns quantitative analysis with an understanding of the artistic context, bridging the gap between computational techniques and the appreciation of language.

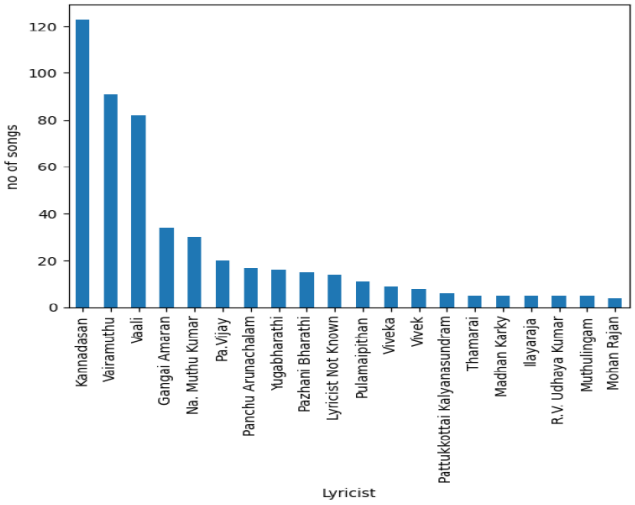

Figure 2 Number of songs written by lyricist with non-Code-mix words (Top 20)

The Tamil lyrics corpus, spanning from 1954 to 2019 shown in Figure 2, comprises 5449 songs, representing various themes and genres. An intriguing observation is the presence of 588 songs written exclusively in pure Tamil, devoid of code-mixing. Among the lyricists, Kannadasan stands out for his linguistic purity, writing the highest number of songs in pure Tamil. His work reflects traditional Tamil literary values. Vairamuthu and Vali also contribute significantly with pure Tamil lyrics, reflecting their mastery of the language. Various other lyricists have also contributed to this corpus, showcasing a blend of tradition and modernity.

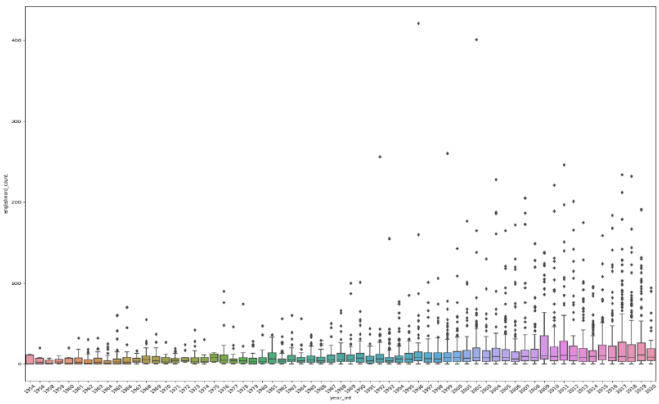

Figure 3 Box plot showing the English word usage per year.

A consistent increase in the use of English words in Tamil songs over the last two decades suggests a shift in lyric-writing culture. This trend, depicted in Figure 3, is likely influenced by the global reach of English music, social media, streaming platforms, and international collaborations. The correlation matrix in Figure 4 shows a strong positive correlation between English words and Tamil code-mix suffix words, indicating a characteristic blending of the two languages.

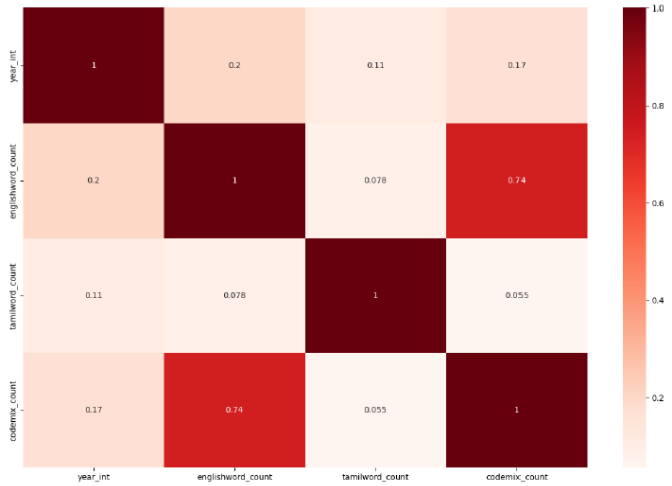

Figure 4 Correlation Matrix

Codemixing is prevalent in multilingual societies, and the high correlation observed may reflect the linguistic tendencies of the population from which the data was sourced. Analysing the linguistic patterns of the top 20 lyricists in terms of English word usage reveals a peak between 2010 and 2020 as shown in Figure 5, reflecting the influence of English media, education, and technology. The integration of Tamil words as suffixes has become a distinctive stylistic choice, resonating with Tamil-speaking audiences and introducing elements of the language to non-Tamil speakers.

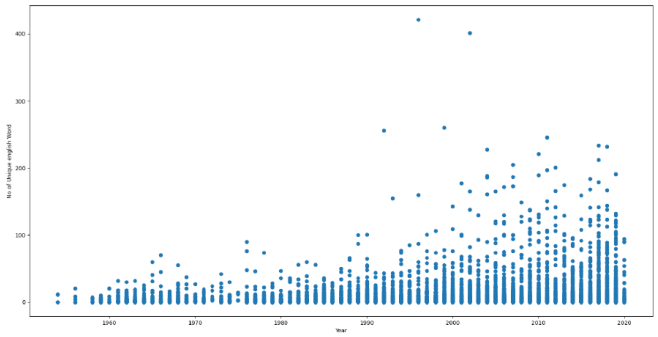

Figure 5 Top 20 lyricist with number of English words used

The increase in Tamil code-mix suffix word usage over the last three decades reflects various influences, as previously discussed. The integration of Tamil suffixes has become a distinctive stylistic choice among lyricists and musicians. Figure 6's bar plot shows the annual counts of these suffixes, revealing that 3-5 Tamil code-mix suffix words have been used in each song released annually since 1990. This trend resonates with Tamil-speaking audiences, providing familiarity and connection, and introduces non-Tamil speakers to elements of the language.

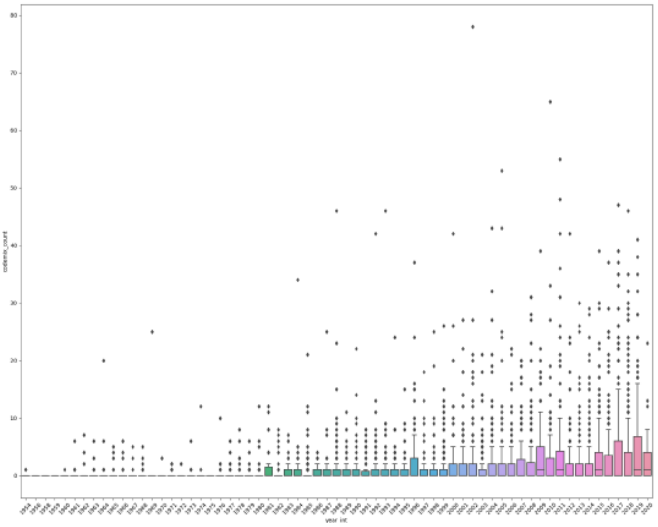

Figure 6 Tamil code-mix suffix word usage in lyrics corpus

Figure 7's box plot shows that an average of 120-160 Tamil words are used per song, indicating a consistent pattern in Tamil songwriting. However, there has been a slight reduction in average Tamil word usage by poets over the last two decades, likely due to the integration of words from other languages

and changing audience interests towards new musical genres.

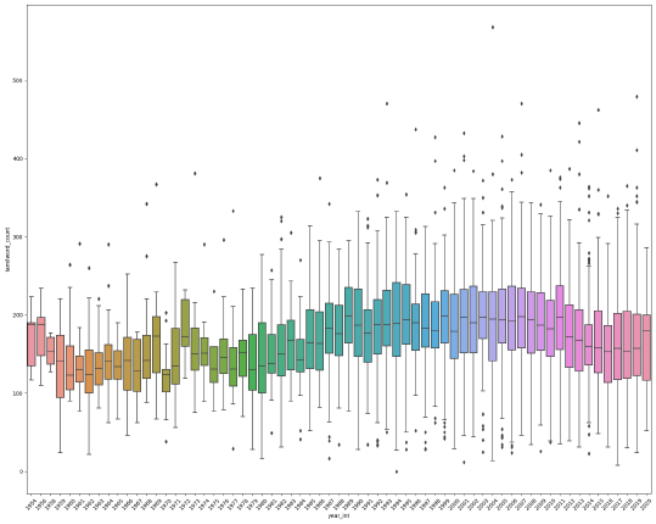

Figure 7 Tamil words usage in lyrics corpus

4. Conclusions

The initial research stages provided a deep understanding of the dataset, highlighting the interplay of English and Tamil words in songs. Kannadasan, a stalwart of traditional Tamil literary norms, wrote most of his songs in pure Tamil without code-mixing. The CRF model, meticulously designed and tested, demonstrated high performance in identifying code-mixed words with precision, recall, F1-score, and accuracy of 0.99. Validation confirmed the model's robustness and ability to generalise to unseen data. K-means clustering revealed well-defined patterns among lyricists' use of code-mixed words.

Notable lyricists like Kannadasan, Vairamuthu, and Vaali were central to this exploration. Vaali and Vairamuthu, using over 2500 unique English words and more than 500 Tamil code-mix suffixes, emerged as top contributors. Next-generation lyricists like Na. Muthu Kumar and Madhan Karky also used significant numbers of English and Tamil code-mix words.

The clustering revealed a hierarchical structure, with Vaali and Vairamuthu at the top, followed by lyricists like Na. Muthu Kumar and Madhan Karky, and others like Kannadasan and Gangai Amaran in subsequent clusters. The least contributing cluster included lyricists like Mohan Rajan and Hiphop Tamizha.

Analysis showed a distinct increase in English and Tamil code-mix words after 1990. Statistical validation with T-Tests and Z-Tests confirmed a significant difference in code-mix word usage between periods, confidently rejecting the null hypothesis.

References

[1] Benhur, S. and Sivanraju, K., (2021) Pretrained Transformers for Offensive Language Identification in Tanglish. CEUR Workshop Proceedings, 3159, pp.659–666.

[2] Chakravarthi, B.R., Muralidaran, V., Priyadharshini, R. and McCrae, J.P., (2020) Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text. [online] Available at: http://arxiv.org/abs/2006.00206.

[3] Gundapu, S. and Mamidi, R., (2018) Word level language identification in english telugu code mixed data. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, PACLIC 2018*, pp.180–186.

[4] Sinaga, K.P. and Yang, M.S., (2020) Unsupervised K-means clustering algorithm. *IEEE Access*, 8, pp.80716–80727.