Tamil lyrics corpus: Analysis and Experiments

Dhivya Chinnappa
Thomson Reuters
dhivya.infant@gmail.com

Praveenraj Dhandapani
Intellect Design Arena Ltd.
praveenraj0904@gmail.com

Abstract

In this paper, we present a new Tamil lyrics corpus extracted from Tamil movies captured across a range of 65 years (1954 to 2019). We present a detailed corpus analysis showing the nature of Tamil lyrics with respect to lyricists and the year which it was written. We also present similarity score across different lyricists based on their song lyrics. We present experimental results based on the SOTA BERT Tamil models to identify the lyricists of a song. Finally, we present future research directions encouraging researchers to pursue Tamil NLP research.

1 Introduction

Tamil is a classical Dravidian language spoken in the states Tamil Nadu and Pondicherry, India. It is also widely spoken in Srilanka and is one of the official languages of the country. There are considerable Tamil speaking people in countries like Singapore, Malaysia, Canada, and Mauritius.

Historians note that Tamil is as old as 2600 years. The earliest identified Tamil literature known as the *Sangam literature* dates back to 600 B.C. (Abraham, 2003). Tamilians as an ethnic group appreciate literary works and refer to literary excerpts in their daily lives. For example, *Thirukkural*[1], one of the popular literary works, also called the *Ulaga Pothu Marai* (translation: Universal book of principles) includes short literary phrases emphasizing on principles for a prosperous life. Schools teach *Thirukkural* to their students as early as kindergarten. *Thirukkural* excerpts are often found in public transportations and government buildings in Tamil Nadu, India. Political leaders, orators and even common people quote *Thirukkural* excerpts in their daily lives.

Since the inception of the Tamil movie industry in the early 1930s, Tamil literary works have taken a different form. From the very beginning, Tamil movies used the platform to showcase the language's rich literature through movie dialogues and song lyrics. Tamil literary works as old as *Sangam literature* have been depicted in early movies, and the recent movies showcase a mix of old and contemporary literary works. Thus, Tamil movies have become an inseparable part of most Tamilians' lives. While independent literary works (poems, hykoos, novels, etc.) exist, Tamil movies consistently act as important medium in showcasing rich Tamil literary works.

The ideologies portrayed in Tamil movies closely reflect Tamilians' lifestyles. Tamil movies and Tamil politics go hand in hand (Kalorth, 2021). There are several people in the Tamil movie industry with strong political opinions, and it is not uncommon for Tamil movie dialogues and Tamil songs to reflect their political ideologies.

The Tamil movie industry releases a movie every four days (Krishnan and Sakkthivel, 2010). That is, on average there are 90 movies released every year. Similar to other Indian language movies, Tamil movies include several songs. Each song in a Tamil movie is written by a lyricist coordinating with the movie director and the music director. The movie director usually describes the situation and the emotion where the song is supposed to fit in. Depending on the given situation by the movie director, the lyricist writes the song lyrics, and the music director composes the music for the

[1] https://thirukkural133.wordpress.com/contents/

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 1–9

song. Though the lyrics of a Tamil movie song can stand alone, they are usually an integral part of the movie enhancing an emotion.

In this paper, we introduce a new Tamil lyrics corpus[2]. Specifically, we focus only on Tamil movie lyrics and not Tamil lyrics by independent lyricists. The main contributions of this paper are (a) a corpus including Tamil song lyrics, lyricists, and other details; (b) a detailed corpus analysis focusing on Tamil lyrics; (c) experimental results to identify the lyricist of a given song lyric; and (d) future research directions about how to utilize the Tamil lyrics dataset.

2 Background

Songs are an integral part of Tamil movies. They represent, enhance, and complement an emotion in the movie. We explain the role of these songs and exemplify with some recent Tamil movie songs.

One of the most common emotion expressed in Tamil songs is romance. This includes *falling in love*, *breakup*, *marriage*, etc. For example, the song *Nenjukul peidhidum*[3] from the movie *Vaaranam Aayiram* describes a situation when a man meets a woman for the first time. This song emphasizes the emotion of *love at first sight*. The visualization and the lyrics are dramatic, and do not depict the actual story of the movie. While the flow of the movie is not affected by the song, it plays a vital role in emphasizing a man's emotion towards a woman the first time he sees her.

Some Tamil movie songs are created to complement the scenes in the movie. The movie would not be complete without the song. These songs include sequential scenes from the movie. For instance, the song *Kadhaipoma*[4] from the movie *Oh my kadavulae* describes the emotion of *taking someone for granted* while aptly fitting in the movie scenes.

There are several Tamil movie songs written with political and social intentions. Tamil actors with political ideologies often hint their political intentions in their songs. Recently, there are several songs released emphasizing on social equity. The song *Aalaporan Thamizhan*[5] from the movie *Mersal* describes the beauty of the language Tamil and glorifies Tamilians life styles.

Recently with the advent of youtube, there are several songs that get popular even before the movie release. While releasing songs before releasing a movie has been a practice in the Tamil movie industry for a while, youtube has taken these songs sooner to the audience than ever before. Until recently, all movie songs were released together in a Compact Disk, cassette tape or an online streaming site. However, in the past few years Tamil movies have started releasing singles from movies. These songs typically include a catchy lyric, music or dance moves. They act as a promotion for the movie and usually have millions of views before the movie release. Kutty story[6] from the movie *Master* and *Bujji*[7] from the movie *Jagamae thanthiram* fall under this category. Interestingly, the song *Kutty story* has large amount of English songs, which is a phenomena observed in many recent Tamil songs.

We present youtube links in the footnote to each song exemplified in this section. The English translation of the song lyrics can be viewed by turning on the subtitles in the youtube video.

How are Tamil song lyrics written? In most cases, the movie director comes up with a situation in the movie.. The music director composes a tune capturing the emotion of this situation. The lyricist writes the lyrics, given the tune. However, there are few music directors who compose tune to a given lyric.

3 Related work

There exists a Tamil lyrics dataset (Subramanian, 2020) in Kaggle for Tamil lyrics. The dataset includes the lyrics of Tamil songs, alongside song name and movie name. However, we could not track any further experiments or research related to this dataset. It does not include the name of the lyricist or other details. Unlike this dataset, our corpus includes the name of the lyricist, the name of the music director, and the year the song was released. We also include the English translit-

[2] https://github.com/praveenraj0904/tamillyricscorpus
[3] Nenjukul peithidum
[4] Kadhaipoma
[5] Aalaporan Thamizhan
[6] Kutty Story
[7] Bujji

erated song lyrics. We believe these information would be useful to researchers.

There are some prominent Tamil NLP works associated with Tamil lyrics. Sundara Kanchana and Ganapathy (2017) work with 1, 000 Tamil songs to identify three aspects for a given Tamil song lyric. The three aspects are base, mood, and style. The base aspect includes the types *character, festival, nature, relationship, romance, occasion, spiritual, patriotic and misc*. The emotion aspect includes the types *happy, sad, excited, tender, scared and angry*. The styles aspect includes *traditional, folk, contemporary and mixed*. Their work is word based, and they use TF-IDF as a feature in their classifier to predict these aspects.

Ramakrishnan A et al. (2009) work on automatically generating Tamil lyrics for melodies. They approach the problem as a sequence labeling problem targeting to guess the syllabic pattern. Then they forward this pattern to a sentence generation module, where they use Dijkstra's algorithm to choose a meaningful phrase. In their latter work Ramakrishnan A and Lalitha Devi (2010), follow a mapping scheme for matching melody with words. They also use a knowledge-based text generation algorithm on an existing Ontology and Tamil Morphology Generator.

E. et al. (2011) design a pleasantness scoring model based not only based on phonetic aspects. Ranganathan et al. (2011) present analysis over a given set of lyrics. In their later work, they (Ranganathan et al., 2013) present a visualization tool for lyrics based on the characteristics of the lyrics.

Sridhar et al. (2013) have extensively worked in automatic Tamil lyrics generation. In their first work, (Sridhar et al., 2013) they follow a trigram approach using the emotion of a given scene. The use this emotion as a seed word and follow Tamil grammatical rules to generate lyrics. In their next work, Sridhar et al. (2014) they follow a similar approach getting 77% accuracy. Following that, they (Sridhar et al., 2015) intend to generate Tamil lyrics given a sequence of images and a derived tune. They generate a tune automatically based on carnatic music. Their system generates two sets of Tamil lyrics based on (i) context text and tune; and (ii) context text and images. In their next work (Sridhar et al., 2016) they follow a neuro-linguistic approach for lyric generation intending to add dialect, image, and author specific features. In their recent work (Sridhar et al., 2018) they generate lyrics, combining their two previous works. They make use of visual images, derived tune, and other specification such as the dialect of the speaker.

Unlike the works described, our work tries to capture the characteristics of the lyric and the lyricist. In an analysis perspective, our work is analogous to (Ranganathan et al., 2011). We go beyond, and present analysis with respect to the year the lyrics were written, and similarity between pairs of lyricists.

Statistics	
Songs	5,449
Lyricists	324
Music directors	151
Movies	1,147
Words	1,309,425
Unique words	129,614
Years range	1954-2019
Maximum words/song	653
Minimum words/song	17
Average words/song	188

Table 1: Statistics of the Tamil lyrics dataset

4 Corpus creation

We create a corpus of Tamil lyrics from www.allnewlyrics.com. The corpus consists of 5,449 Tamil songs lyrics and their English transliteration. Apart from the lyrics of a song, the corpus includes the name of the lyricist, music director, movie, and the year the song was released. We also present a link to the web page where the data was extracted. The statistics of the corpus are presented in Table 1. The corpus includes songs written by 324 lyricists across 1,147 movies for which the music was composed by 151 music directors.

From a linguistics perspective, lyrics are the most interesting. The lyrics of a song reflects the (i) situation in the movie, (ii) writing style of the lyricist, (iii) language style during which the lyrics were released, and (iv) the emotion of the song. On an average each song includes 188 words ranging from minimum 17 words to

	Lyricist	Music director	Movie	Year
	Vaali	M.S. Vishwanathan	Kasethan Kadavulada	1972
Tamil	மெல்ல பேசுங்கள் பிறர் கேட்க கூடாது சொல்லித் தாருங்கள் யாரும் பார்க்க கூடாது [...]			
English	Mella pesungal pirar ketkka koodadhu solli thaarungal yaarum paarkka koodadhu [...]			
	Lyricist	Music director	Movie	Year
	Vairamuthu	A. R. Rahman	Love Birds	1996
Tamil	மலர்களே மலர்களே இது என்ன கனவா மலைகளே மலைகளே இது என்ன நினைவா [...]			
English	Malargalae malargalae idhu enna kanavaa malaigalae malaigalae idhu enna ninaivaa aa... [...]			
	Lyricist	Music director	Movie	Year
	Na. Muthu Kumar	Yuvan Shankar Raja	Kaadhal kondein	2003
Tamil	நெஞ்சோடு கலந்திடு உறவாலே காலங்கள் மறந்திடு அன்பே நிலவோடு தென்றலும் வரும் வேளை காயங்கள் மறந்திடு அன்பே [...]			
English	Nenjodu kalanthidu uravaalae kaalangal maranthidu anbae nilavodu thendralum varum velai kaayangal maranthidu anbae [...]			
	Lyricist	Music director	Movie	Year
	Gangai Amaran	Ilayaraja	Paneer Pushpangal	1981
Tamil	கோடைக்கால காற்றே குளிர்த்தென்றல் பாடும் பாட்டே மனம் தேடும் சுவையோடு தினம்தோறும் இசைபாடு [...]			
English	Kodai kaala kaatrae kulir thendral paadum paattae manam thedum suvaiyodu dhinandhorum isai paadu [...]			
	Lyricist	Music director	Movie	Year
	Thamarai	Darbuka Siva	Enai Noki Paayum Thota	2019
Tamil	எதுவரை போகலாம் என்று நீ சொல்லவேண்டும் என்றுதான் விடாமல் கேட்கிறேன் [...]			
English	Ethuvarai pogalaam endru nee sollavendum endruthaan... vidamal ketkiren... [...]			

Table 2: Examples from the Tamil lyrics corpus. Each instance includes the name of the lyricist, music director, movie, and the year the movie was released. Additionally, the Tamil lyrics and the equivalent English transliterated lyrics is also available.

maximum 653 words. The total number of Tamil words for all songs combined in the corpus is more than a million. The total number of unique Tamil words in the corpus is close to 130, 000. Table 2 presents examples from the corpus.

5 Corpus Analysis

We present detailed corpus analysis focusing on the Tamil lyrics our corpus. All statistics presented in this paper represent the distribution of the data in our corpus and may not reflect the contribution of a lyricist to the Tamil movie or music industry[8].

Figure 1 depicts the number of songs written by a lyricist across different years. Evidently, Vaali has written songs across a wide range of years starting from the mod 1960s to the late 2010s. Vairamuthu has reached the peak of his lyric writing in the early 2000s. Interest-

[8]All statistics correspond to the corpus, and may not reflect the contribution of a lyricist in real time.

Lyricist	No. songs	Top 10 words
All (with stopwords)	5449	நீ, என், நான், உன், ஒரு, என்ன, காதல், வா, தான், என்னை
All (No stopwords)	5449	தா, போது, மேலே, பார்த்து, பொண்ணு, இவன், உலகம், தேன், கண்ணே, சுகம்
Vaali	873	தா, வந்தது, நாளும், தேன், மாமா, தொட்டு, மெல்ல, பிள்ளை, வாழும், அவன்
Kannadasan	797	கண்டேன், உள்ளம், கொண்ட, வந்தது, அவள், ஆட, கண்ணா, பிள்ளை, கண்டு, ஒரே
Vairamuthu	390	நோ, முத்தம், கண்ணே, கண்டு, ஆண், உலகம், கிளி, பூமி, சிரி, சத்தம்
Na. Muthu Kumar	340	இதயம், உயிரே, இவன், உலகம், ராஜா, அவள், பறக்குது, எங்கும், வேணும், முன்னே
Gangai Amaran	279	மாமா, ஊரு, பொண்ணு, கேட்டு, குயிலே, கேளு, கதை, ஒண்ணு, ராகம், ராசா
Pa.Vijay	263	பேபி, முதல், நோ, போடு, வாவ், சும்மா, தானா, தில், முனி, தீண்ட
Yugabharathi	179	ஜிக்கி, அழகு, புள்ள, அவ, போ, பேச, கண்ணு, ஆச, வாழ, அப்பா
Madhan Karky	163	வெறி, மை, ராஜா, ராஜ, அவள், சச்சின், மேலே, தல, ஏத்திக்க, நாம்
Thamarai	107	ஏனோ, யாரோ, அவள், முதல், நிலா, என, எனை, மேகம், மேலே, தாண்டி
Viveka	105	ரக்கரா, கபடி, சிக்கு, டக்குனக்கா, மட்டக்கு, கோ, கைய, ஒய்ய, மியாவ், சக்கான்

Table 3: Number of songs written by a lyricist and the most frequent 10 words by them are presented. The most frequent 10 words are obtained after removing the stop words. The most frequent 100 words from the entire corpus are treated as stop words.

ingly, Na. Muthukumar has written the most number of songs in the shortest span of years.

We treat each word in the corpus as the simplest unit for our analysis. That is, we do not conduct analysis based on lemmas or stems of words. The top 10 lyricists with the highest number of songs and their most frequent 10 words are presented in Table 3. Vaali ranks first with 873 songs followed by Kannadasan (797 songs), and Vairamuthu (390 songs).

We noticed that the top 10 most frequent words across the entire corpus (Table:3, 1st row), and the top 10 most frequent words per lyricist were all pronouns or interrogative words. Our further investigation revealed that the top 100 words mostly included pronouns, interrogative words, prepositions, and conjunctions which are considered as stop words in English. Thus we decided to treat the top 100 most frequent words as stop words. We acknowledge that this process is not perfect, as there are a few non stop words in the top 100 words, and few stop words after the most frequent 100 words. The top 10 most frequent words per lyricist (Table 3) are extracted after removing the stop words. Interestingly, the word காதல் (love) always is present in the top 10 words despite not being a traditional stop word. We attribute this to the nature of Tamil movie songs which usually circles around the emotion, romance or love (காதல்).

Table 4 presents the number of songs per decade and the top 10 most frequent words per decade. For example, there are 312 songs included in the corpus belonging to the decade (1970), that is from the year 1961 to 1970. The top 10 words for the decade 1970 are also given. The top 10 most frequent words are extracted after removing the stop words.

In Table 5, we present similarity scores between pairs of lyrics among the top 5 artists. The Jaccard similarity computes similarity between two sets by calculating the intersection over union. First, we combine all songs by

Year	No. songs	Top 10 words
1960	71	கண்டேன், கனவு, எங்கள், சிவகங்கை, சீமை, சொந்தம், வாழும், அப்பா, மொழி, சுகம்
1970	312	உள்ளம், அவள், கொண்ட, பார்த்து, நிலா, ஆட, அவன், பார், கண்டு, கண்டேன்
1980	296	சுகம், பிள்ளை, வந்தது, ஆடும், இரு, கண்ணன், உள்ளம், மெல்ல, ராகம், பா
1990	820	ராகம், கதை, மானே, மாமா, தேன், தேகம், நாளும், பாட, வந்தது, சுகம்
2000	1071	நோ, பூவே, தா, பொண்ணு, நிலா, கிளி, கண்ணே, தேன், மாமா, ஒன்னு
2010	1087	முதல், வாவ், போடு, ஏதோ, தொட்டு, இதயம், போது, பார்த்து, மேலே, மெல்ல
2020	1593	இவன், போ, நாங்க, இதயம், அவ, வாடி, ராஜா, அவன், நாம், போக

Table 4: Number of songs written in a decade and the most frequent 10 words per decade is given. The most frequent 10 words are obtained after removing the stop words. The most frequent 100 words from the entire corpus are treated as stop words.

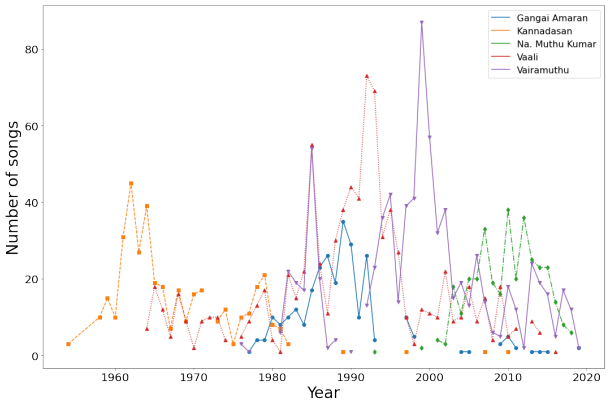

Figure 1: The number of songs written by a lyricist across 65 years (1954 to 2019) is presented. Vaali has written Tamil song lyrics covering the widest range of years. Na. Muthukumar has written the most number of songs in the shortest span of years.

		Jacc. sim.	TFIDF cos sim
Vaali	Kannadasan	0.76	0.956
	Vairamuthu	0.94	0.972
	Na. Muthu Kumar	0.89	0.956
	Gangai Amaran	0.76	0.969
Kannadasan	Vairamuthu	0.71	0.948
	Na. Muthu Kumar	0.77	0.910
	Gangai Amaran	0.85	0.922
Vairamuthu	Na. Muthu Kumar	0.9	0.964
	Gangai Amaran	0.74	0.939
Na. Muthu Kumar	Gangai Amaran	0.81	0.924

Table 5: Words based similarity between pairs of the top 5 lyricists from the corpus. Both Jaccard similarity and cosine similarity on the TF-IDF vectors indicate that the lyrics of Vaali and Vairamuthu are the most similar.

each lyricist into a large text. Thus we obtain five large texts for the top 5 lyricists. We remove duplicate words from each text to create a set for each lyricist. We calculate the Jaccard similarity over these sets to measure their similarity. Results show that there is higher overlap between the words used by Vaali and Vairamuthu (Jacc. sim.: 0.94), and Vairamuthu and Na. Muthu Kumar (Jacc. sim.: 0.9)

We also present the cosine similarity between the TF-IDF representations of the lyrics by the top 5 lyricists. We process the lyrics similar to the processing for calculating Jaccard similarity, but do not form a unique set of words for each lyricist. First, we combine all songs by each lyricist into a large text. Next, we represent these texts using their TF-IDF vector. We calculate the cosine similarity between these vectors to measure their similarity. Based on this TF-IDF cosine similarity measure, the lyrics Vaali and Vairamuthu (TF-IDF cos. sim: 0.972) are very similar which is observed in the Jaccard similarity score (Jacc. sim.: 0.94) as well. Similarly Vaali and Gangai Amaran (TFIDF cos. sim: 0.969), and

	Multingual BERT uncased			Indic-NLP AI4Bharat			Tamillion BERT		
	P	R	F1	P	R	F1	P	R	F1
Vaali	0.62	0.57	0.59	0.65	0.75	0.69	0.68	0.74	0.70
Kannadasan	0.56	0.62	0.59	0.67	0.55	0.60	0.68	0.61	0.64
W.Avg	0.59	0.59	0.59	0.66	0.66	0.65	0.68	0.68	0.68

Table 6: Results of binary classifiers to classify between Vaali and Kannadasan given a song Tamil lyric is presented. All models are fine tuned to obtain the results.

Vairamuthu and Na. Muthu Kumar (TFIDF cos. sim: 0.964). also have high TFIDF cosine similarity scores.

6 Experiments and Results

We conduct experiments to identify the lyricist of a given lyric. We formulate a binary classification problem by generating a dataset including only the top two lyricists with the most number of songs, Vaali and Kannadasan. This dataset consists of 1,670 instances from the Tamil lyrics corpus, where 873 instances correspond to Vaali and 797 instances correspond to Kannadasan. We used 80% of this dataset for training and withheld the remaining 20% for testing. We ensured the lyricists distribution is stratified across the train and test sets.

6.1 Experiments

In an era where BERT (Devlin et al., 2019) models are used as a baseline, we present results using three BERT models supporting Tamil. We conducted all our experiments using the simpletransformers (Rajapakse, 2019) python package. First, we fine tuned the multilingual uncased BERT (Reimers and Gurevych, 2020) model. We obtained the best results at a learning rate of $1e^{-5}$ with 3 epochs. Next, we fine tuned with the indic-bert model from AI4Bharat (Kunchukuttan et al., 2020). Here we obtained the best results at a learning rate of $1e^{-5}$ with 2 epochs. Finally, we fine tuned the Tamillion BERT (Doiron, 2020) model. We reached the best results in this model at learning rate of $1e^{-5}$ with 7 epochs.

6.2 Results

We present results in Table 6. The results obtained using multilingual BERT is lower than Indic-NLP and Tamillion BERT (F1: .59 vs. .65 vs. .68). The results obtained using Indic-NLP and Tamillion BERT are comparable (F1: .65 vs. .68).

Regarding the results for *Vaali*, Tamillion BERT obtains the highest precision (.68), and Indic-NLP obtains the highest recall (.75). Overall, the performance for *Vaali* is better when using the Tamillion BERT (F1: .70).

In case of the results for *Kannadasan*, the highest performance is obtained both for precision and recall by fine tuning the Tamillion BERT model (P: .68, R: .61). We note that there is still room for improvement in this binary lyricist identification task. The task becomes even more challenging when the number of lyricists are increased changing the binary classification task into a multi class classification task.

7 Future research

We use the Tamil lyrics corpus to identify lyricists formulating a binary classification task. The same corpus could be used for several other tasks. We discuss the previous works relating to Tamil lyrics in Section 3. In this section, we describe a few possible research that could conducted using the Tamil lyrics corpus.

7.1 Language trend analysis

Language trend analysis intends to capture the change in a language over years. As Tamil song lyrics are closely tied to the political and personal lives of Tamilians, we expect the change of language style in the lyrics to reflect the lives of Tamilians. We hypothesize that the songs released after the year 2000 would include more English words (code-switching) than the songs from the 1960s. For instance, the song *Why this kolaveri?* has more than 50% English words. Additionally, as Tamilians have moved across the world, Tamil songs

have gained international attention, enabling changes in Tamil music.

We also hypothesize that the songs released after the year 2000 would include several meaningless words, whereas the songs released around the 1960s followed strict literary style. For example, phrases like jimjikka, rangu rakkara, tasaku tasaku, etc. have no meaning (also called *Rettai Kilavi* in Tamil grammar) are often used in Tamil songs, as they are fun and catchy.

7.2 Emotion detection

As Tamil song lyrics are usually written for an emotion in the movie, it is possible to use the Tamil lyrics dataset to identify the emotion in them. This task would require an additional layer of annotations for emotions. We also believe a distant supervision mechanism to annotate for the labels should be possible. The possible emotions would include, but not limit to *love, breakup, hero introduction, happy,* and *sorrow*. While the work by (Sundara Kanchana and Ganapathy, 2017) closely relates to the emotion detection space, we believe there is so much to be explored in this area.

7.3 Evaluating transliteration models

There are many works relating to English transliteration of Tamil (Chinnakotla and Damani, 2009; Vijayanand, 2009). As our corpus includes the English transliterated text of the Tamil lyrics, it could be used as a validation set to evaluate a transliteration model.

We hypothesize that Tamil lyrics are rich in linguistic phenomena such as cliches, metaphors, etc. A model identifying such phenomena in the Tamil corpus would be interesting.

8 Conclusion

In this paper, we introduce a new Tamil lyrics dataset. We collect Tamil movie lyrics along with the names of the lyricist, the movie, and the music director over a range of 65 years. We also present the English transliteration of the Tamil lyrics. We present an elaborate literature review on studies relating to Tamil lyrics. We conduct detailed corpus analysis revealing the top 10 frequent words per decade, and top 10 frequent words per lyricist. We also study the similarity of the lyrics between different lyricists using standard similarity approaches. Next, we formulate a binary classification task to identify the lyricist, given a song lyric. We fine tuned the state-of-the-art BERT models to predict the lyricists. Our experimental results show that identifying lyricists is a challenging task, and there is scope for improvement. We go beyond the experiments, and elaborate on the possible future research directions with the Tamil lyrics corpus. We intend to add more instances to our corpus to encourage research on Tamil lyrics.

References

Shinu Anna Abraham. 2003. Chera, chola, pandya: Using archaeological evidence to identify the tamil kingdoms of early historic south india. In *Asian Perspectives*, volume 42, pages 207–223. University of Hawai'i Press.

Manoj Kumar Chinnakotla and Om P. Damani. 2009. Experiences with English-Hindi, English-Tamil and English-Kannada transliteration tasks at NEWS 2009. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 44–47, Suntec, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nick Doiron. 2020. Tamillion bert. `https://huggingface.co/monsoon-nlp/tamillion`.

Giruba Beulah S. E., Madhan Karky V., Karthika Ranganathan, and Suriyah M. 2011. Pleasantness scoring models for tamil lyrics. In *Proceedings of the 5th Indian International Conference on Artificial Intelligence, IICAI 2011, Tumkur, Karnataka State, India, December 14-16, 2011*, pages 1561–1571. IICAI.

Nithin Kalorth. 2021. Identities and ideologies in tamil cinema understanding within the framwork of socio political movements in tamil nadu. In *Into the Nuances of Cultural - Essays on Cultural Studies*, pages 95–103. Yking books.

Trichy V. Krishnan and A.M. Sakkthivel. 2010. To push for stardom or not: A rookie's dilemma in

the tamil movie industry. *IIMB Management Review*, 22(3):80–92.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

T. C. Rajapakse. 2019. Simple transformers. https://github.com/ThilinaRajapakse/simpletransformers.

Ananth Ramakrishnan A, Sankar Kuppan, and Sobha Lalitha Devi. 2009. Automatic generation of tamil lyrics for melodies. In *Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity*, pages 40–46, Boulder, Colorado. Association for Computational Linguistics.

Ananth Ramakrishnan A and Sobha Lalitha Devi. 2010. An alternate approach towards meaningful lyric generation in Tamil. In *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*, pages 31–39, Los Angeles, California. Association for Computational Linguistics.

K. Ranganathan, T. V. Geetha, R. Parthasarathi, and Madhan Karky. 2011. Lyric mining : Word , rhyme & concept co-occurrence analysis.

Karthika Ranganathan, B. Barani, and T. V. Geetha. 2013. A tamil lyrics search and visualization system. volume 8281, pages 513–527. Lecture Notes in Computer Science.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

R. Sridhar, N. Dharmaraj, J. Damodaran, and S. Sampath. 2015. Automatic tamil lyric generation based on image sequence and derived tune. In *2015 IEEE International Advance Computing Conference (IACC)*, pages 861–866.

R. Sridhar, S. Venkatsubhramaniyen, and S. S. Rashmi. 2018. Automatic singable tamil lyric generation for a situation and tune based on causal effect. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1152–1159.

Rajeswari Sridhar, S. R. Surya Dev, V. Sankar, and K. Sivakumar. 2016. Automatic tamil lyric generation based on neuro-linguistics. In *Proceedings of the International Conference on Informatics and Analytics*, ICIA-16, New York, NY, USA. Association for Computing Machinery.

Rajeswari Sridhar, G Jalin Gladis, K Ganga, and G Dhivya Prabha. 2013. N-gram based approach to automatic tamil lyric generation by identifying emotion. pages 919–926. Proceedings of International Conference on Advances in Computing.

Rajeswari Sridhar, G Jalin Gladis, K Ganga, and G Dhivya Prabha. 2014. Automatic tamil lyric generation based on ontological interpretation for semantics. pages 97–121. Academy Proceedings in Engineering Sciences.

Siva Subramanian. 2020. Tamil songs lyrics dataset. Data retrieved from Kaggle http://www.kaggle.com/sivaskvs/tamil-songs-lyrics-dataset.

Meenakshi K Sundara Kanchana and Velappa Ganapathy. 2017. Comparison of genre based tamil songs classification using term frequency and inverse document frequency. *Research Journal of Pharmacy and Technology*, 5:80–92.

Kommaluri Vijayanand. 2009. Testing and performance evaluation of machine transliteration system for Tamil language. In *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009)*, pages 48–51, Suntec, Singapore. Association for Computational Linguistics.