MDOLC: Multi Dialect Odia Song Lyric Corpus

Conference Paper · February 2023

CITATIONS
0

READS
1,490

4 authors:

Shantipriya Parida
Silo AI
106 PUBLICATIONS **597** CITATIONS
SEE PROFILE

Alakananda Tripathy
Utkal University
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Satya Ranjan Dash
KIIT University
170 PUBLICATIONS **808** CITATIONS
SEE PROFILE

Shashikanta Sahoo
Biju Patnaik University of Technology
5 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Grenze International Journal of Engineering and Technology, Jan Issue

MDOLC: Multi Dialect Odia Song Lyric Corpus

Shantipriya Parida[1], Alakananda Tripathy[2], Satya Ranjan Dash[3], and Shashikanta Sahoo[4]
[1] Silo AI, Helsinki, Finland
Email: shantipriya.parida@silo.ai
[2]Utkal University, Bhubaneswar, India
Email: alakananda.ranjan@gmail.com
[3]KIIT University, Bhubaneswar, India
Email: sdashfca@kiit.ac.in
[4]Government College of Engineering, Kalahandi, India
Email: shashisahoo123@gmail.com

***Abstract*—This paper presents MDOLC, a multi-dialect Odia song lyric corpus. The corpus contains 230 Odia song lyrics in two different dialects (standard Odia or Odia and Sambalpuri). The lyrics are segmented into more than 8033 sentences (song verses) with 42,287 words. The corpus is available in annotated plain text (txt) format with metadata. We perform the dialect detection experiment on this corpus. Also, we performed a linguistic analysis of this corpus. The MDOLC is the first multi-dialect Odia song lyric corpus as per our best knowledge and will be freely available for researchers.**

***Index Terms*— Lyric Corpus, Dialect, Multilingual, Language Classification.**

I. INTRODUCTION

Odia belongs to the Indo-Aryan language family. Odia is a low-resource language spoken by 37 million speakers[1] in the Odisha state of India. Odia has proven its ancient classical legacy to have more than 1500 years of continuous history due to its rich antiquities, inscriptions, and available scriptures [1]. Odia has been heavily influenced by the Dravidian languages as well as Arabic, Persian, and English [1]. Borrowings have enriched its lexicon from these languages as well as from Tamil, Telugu, Marathi, Turkish, French, Portuguese, and Sanskrit[2].

The lyrical journey of the Odia language traces back to the origin of the Jagannath cult in Odisha [2]. Through theaters, folk songs, folk dance, and folk theaters (Pala, kirtan, bhajan, gitinatya, etc.), mythological-social stage plays have flourished through the centuries. However, cinematic representation of Odia lyrics started with the talkative movie "Sita Bibaha" (Sita's marriage") which was a drama written by Kamapal Mishra and directed by Mohan Sundardev Goswami in 1936 [3].

Odia has many regional dialects like Sambalpuri, Berham-puri, Baleswari, Koraputi, Desia, and Sundargadia[3]. Sambalpuri (ISO 639-3)[4] is the western dialect/variety of Odia language spoken in the western part of Odisha state and has a significant contribution to lyrical songs in Odia language apart from standard Odia [4].

[1] https://censusindia.gov.in/2011Census/Language_MTs.html
[2] https://www.britannica.com/topic/Odia-language
[3] https://glottolog.org/resource/languoid/id/oriy1255
[4] https://www.ethnologue.com/language/spv

Grenze ID: 01.GIJET.9.1.723

In this paper, we collected various categories of Odia songs and Odia songs with Sambalpuri dialects from different song websites to create a corpus of Odia song lyrics. The frequent words in Odia and Sambalpuri dialects are shown in Figure 2. As the collection of songs is modern, it includes words from Hindi and English as well such as the Hindi word “dil (heart)”, and the English word “love” as shown in Figure 1, and Figure 2.

Figure. 1. Word cloud of Odia dialects

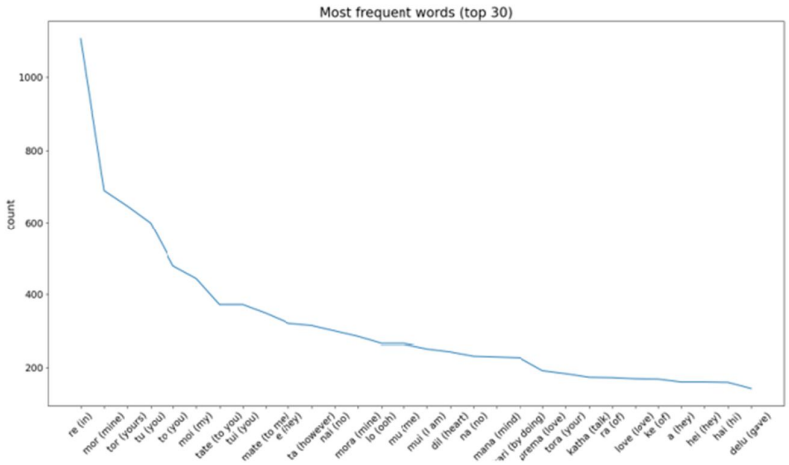

Figure. 2. Most frequent Odia dialect words with English translation (top 30)

II. Related Work

Although there exists corpora of song lyrics in other languages including Indian languages (Hindi, Tamil, Punjabi) for Natural language processing (NLP) research [5]–[10], the availability of song lyric corpus for the low-resource Odia language is limited. There has not been much attention made to compiling Odia songs of different categories and dialects into a corpus suitable for NLP research and our work is one of the first of its kind in this direction. A collection of 730 Odia poems (named “Kabithaa”), annotated and released for NLP research, particularly for sentiment analysis [11] - [12].

III. Data Description

The statistics of the Odia song categories are shown in Table I.

TABLE I. SONG CATEGORIES

#	Songs categories	Count
1	Odia modern songs	142
2	Odia songs with Sambalpuri dialect	88

IV. DATA PROCESSING

We collected Odia songs mainly from the Odia song lyrics websites (see Table II) and preprocessed them for the experiment and release the corpus. The collection of song lyrics data and metadata from the web (e.g. lyric wiki, lyrics websites) was already followed by many researchers in multiple languages which we followed [5], [13], [14]. The preprocessing includes:

- Removal of unwanted characters.
- Avoid repetition of sentences.
- Convert to lowercase.

TABLE II. SONG LYRIC SOURCES

#	Website
1	https://allodialyrics.com/
2	https://odia-lyrics.blogspot.com/
3	https://odialive.com/

We collected the metadata along with the songs and include this data in the corpus as well. In the case of unavailability of metadata, we annotated with "Not Available (NA)" (e.g.<Song Writer>NA< /Song writer>). The metadata collected are shown in Table III and their statistics are shown in Table IV. The corpus is available as annotated plain UTF-8 text files (txt). The plain text (txt) format contains HTML-like annotation tags to identify the song and singer details as well as the lyrics of each song. Figure 3 shows a txt annotated sample, the format is consistent across the corpus with a separate file for each song.

TABLE III. SONG META DATA

1	Song Title
2	Dialect
3	Period
4	Singers
5	Gender of Singers
6	Music Director
7	Gender of Music Director
8	Lyrics Author
9	Gender of Lyrics Author
10	Director
11	Gender of Director
12	Producer
13	Gender of Producer
14	Movie or Album
15	Year

TABLE IV. MDOLC CORPUS STATISTICS

Songs	230
Song Title	230
Sentences	8033
Words	42,287
Singers	90
Dialects	2
Period	1990-2021

V. DIALECT DETECTION

We performed a dialect detection experiment on MDOLC. The reported experiment was conducted at the sentence level where each sentence is a verse from a song, those surrounded with the sentence tags (< s >< /s >) as shown in Figure 3. Each sentence is labeled with the dialect ("ori" for Odia and "spv" for Sambalpuri as per ISO language code). We used both classical and deep learning models for the dialect detection task.

```
<Song ID>1</Song ID>
<Title>mushkil hai jeena</Title>
<Dialect>odia</Dialect>
<Period>after year 1990</Period>
<Singers>babushan & diptirekha</Singers>
<Gender(Singer)>male & female</Gender(Singer)>
<Music Director>prem anand</Music Director>
<Gender(Music Director)>male</Gender(Music Director)>
<Lyrics Author>subrat swain</Lyrics Author>
<Gender(Lyrics Author)>male</Gender(Lyrics Author)>
<Director>ashok pati</Director>
<Gender(Director)>male</Gender(Director)>
<Producer>binni samal & nihar samal</Producer>
<Gender(Producer)>female & male</Gender(Producer)>
<Movie_Album>ajab sanjura gajab love</Movie_Album>
<Year>2017</Year>
<lyrics>
    <s>otha tora madusala</s>
    <s>mitha mitha mahu jhara</s>
    <s>mahu bina prajapati rahi parena</s>
    <s>mushkil hai jeena tere bina</s>
    <s>akhi tora swapana jhara</s>
    <s>tofa tofa janha tara</s>
    <s>janha bina rati jama soi parena</s>
    ...
    ...
    ...

<lyrics>
```

Figure. 3: Odia song corpus sample

VI. EXPERIMENTAL SETUP

As part of the dialect detection experiment, we used the classical Support Vector Classifier (SVC) as a baseline and deep supervised autoencoder (SAE) as a deep learning model.

A. Baseline

As a baseline system, we implemented a binary linear Sup-port Vector Classifier (SVC) using as a form of representation of the documents a traditional bag-of-words (BoW) strategy with a tf-idf weighting scheme.

B. Baseline Deep Supervised Autoencoder

We used the SAE with bayesian optimizer for the dialect detection task which was already found effective in language and dialect detection tasks [15], [16]. We followed the approach by [15].

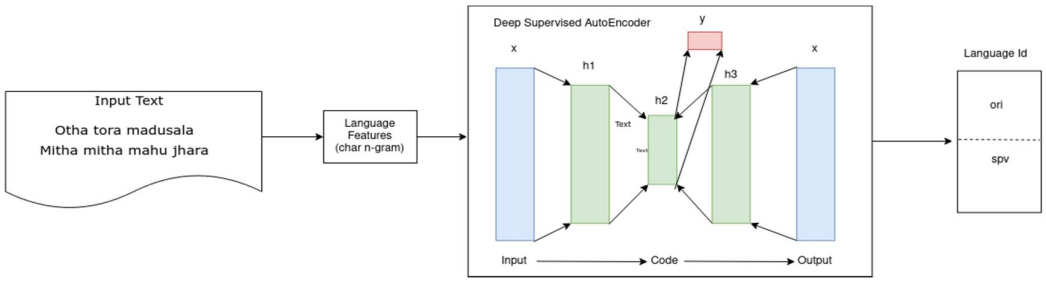

Figure. 4: Proposed model architecture. The extracted features of the text are input to the supervised autoencoder. The target "y" is included. The classification output is the language id for the classified languages (ori: Odia, spv: Sambalpuri).

C. Textual Features

Character n-grams are input to the SAE. In comparison to word n-grams, which only capture the identity of a word and its possible neighbors, character n-grams are better and additionally capable of detecting the morphological makeup of a word [17], [18]. The extracted n-gram features are input to the deep SAE as shown in Figure 4. The deep SAE contains multiple hidden layers. The hyperparameters were optimized using Bayesian Optimizer (BO).

D. Hyperparameters

The range of values for the hyperparameters search space is shown in Table V. During training, BO chooses the best hyperparameters from this range. The overall configuration of the SAE model is shown in Table VI.

TABLE V. SEARCH SPACE HYPERPARAMETER RANGE

Hyperparameter	Range
number of layer	1-5
learning rate	10^{-5} - 10^{-2}
weight decay	10^{-6} - 10^{-3}
activation functions	'relu', 'sigma'

TABLE VI. SAE MODEL CONFIGURATIONS FOR THE DATASET

Parameter	Odia-Sambalpuri
n_gram range	1-3
number of target	2
embedding dimension	300
supervision	'clf'
converge threshold	0.00001
number of epochs	30

E. Datasets

For the experiment, we divided the total number of sentences into an 80:10:10 ratio for the train/dev/test set as shown in Table VII.

TABLE VII. DATASET STATISTICS (NUMBER OF SENTENCES)

Dataset	Training	Development	Test
Odia-Sambalpuri	6426	804	803

VII. RESULTS AND ANALYSIS

The SAE and SVC model's performance in terms of classification accuracy is shown in Table VIII. Both models perform equally on the test set.

TABLE VIII. OVERALL PERFORMANCE OF THE PROPOSED APPROACH

		Accuracy	
Model	Dataset	Dev	Test
SAE (char-3gram)	Odia-Sambalpuri	90%	92%
SVC (tf-idf)	Odia-Sambalpuri	93%	92%

The SAE model performance (Precision, Recall, and F1) score for each class on the test set is shown in Table IX.

TABLE IX. MODEL PERFORMANCE FOR EACH CLASS (PRECISION, RECALL, AND F1) ON TEST SET

Model	Class	Precision	Recall	F1
SAE	Odia (ori)	0.90	0.93	0.91
	Sambalpuri (spv)	0.93	0.90	0.92
SVC	Odia (ori)	0.92	0.91	0.92
	Sambalpuri (spv)	0.92	0.92	0.92

As many words are shared between Odia and Sambalpuri, the classification models (SVC and SAE) failed to classify A few instances as depicted in the confusion matrix in Figure 5 and Figure 6.

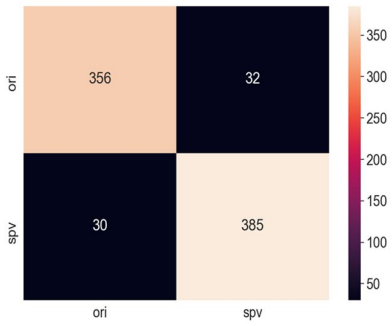

Figure. 5: Confusion matrix (Support Vector Classifier)

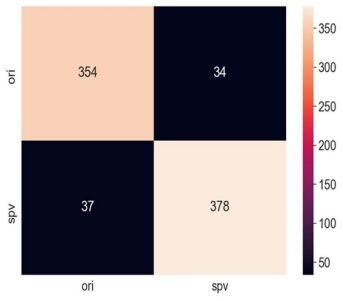

Figure. 6: Confusion matrix (Supervised Autoencoder)

VIII. Availability

After the paper's acceptance, the Odia song lyric corpus will be released for non-commercial research purposes.

IX. Conclusion and future work

In this paper, we propose a multi-dialect Odia song lyric corpus for the low-resource Odia language suitable for NLP research, particularly i) lyric analysis [19], ii) dialect detection el2020habibi, iii) lyric code-mix [20], iv) topic modeling [21], and iv) Sentiment analysis.
The future work includes i) Extending the dataset with more song lyrics of multiple dialects (Berhampuri, Baleswari, Koraputi, and Sundargadia), and ii) experimenting with other similar dialects of Odia for performance evaluation.

References

[1] S. Parida, S. R. Dash, O. Bojar, P. Motlicek, P. Pattnaik, and D. K. Mallick, "Odiencorp 2.0: Odia-english parallel corpus for machine translation," in Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation, 2020, pp. 14–19.

[2] S. C. Jnana and S. Drusti, "The cult of lord jagannath and its impact on oriya literature," ORISSA REVIEW, p. 116.

[3] P. Roy, "Aesthetics of emotional acting: an argument for a rasa-based criticism of indian cinema and television," Ph.D. dissertation, University of Edinburgh, 2017.

[4] P. Behera, A. K. Ojha, and G. N. Jha, "Issues and challenges in developing statistical pos taggers for sambalpuri," in Language and Technology Conference. Springer, 2015, pp. 393–406.

[5] M. El-Haj, "Habibi-a multi dialect multi national arabic song lyrics corpus," 2020.

[6] M. Fell, E. Cabrio, E. Korfed, M. Buffa, and F. Gandon, "Love me, love me, say (and write!) that you love me: Enriching the WASABI song corpus with lyrics annotations," in Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, May 2020, pp. 2138–2147. [Online]. Available: https://aclanthology.org/2020.lrec-1.262

[7] C. Strapparava, R. Mihalcea, and A. Battocchi, "A parallel corpus of music and lyrics annotated with emotions." in LREC. Citeseer, 2012, pp. 2343–2346.

[8] G. D. Apoorva and R. Mamidi, "Bolly: Annotation of sentiment polarity in bollywood lyrics dataset," in International conference of the pacific association for computational linguistics. Springer, 2017, pp. 41–50.

[9] D. Chinnappa and P. Dhandapani, "Tamil lyrics corpus: Analysis and experiments," in Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 1–9.

[10] J. R. Saini and J. Kaur, "Kavi:¯ An annotated corpus of punjabi poetry with emotion detection based on 'navrasa'," Procedia Computer Science, vol. 167, pp. 1220–1229, 2020.

[11] G. Mohanty, P. Mishra, and R. Mamidi, "Kabithaa: An annotated corpus of odia poems with sentiment polarity information," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France, may. European Language Resources Association (ELRA), 2018

[12] "Sad or glad? corpus creation for odia poetry with sentiment polarity information," in Proc. 19th Int. Conf. Comput. Linguistics Intell. Text Process.(CICLing), 2018.

[13] A. Kilgarriff and G. Grefenstette, "Web as corpus," in Proceedings of Corpus Linguistics, vol. 2001, 2001, pp. 342–344.

[14] M. Fell, E. Cabrio, E. Korfed, M. Buffa, and F. Gandon, "Love me, love me, say (and write!) that you love me: Enriching the wasabi song corpus with lyrics annotations," in Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 2138–2147.

[15] S. Parida, E. Villatoro-Tello, S. Kumar, M. Fabien, and P. Motlicek, "Detection of similar languages and dialects using deep supervised autoencoder," in Proceedings of the 17th International Conference on Natural Language Processing (ICON). Indian Institute of Technology Patna, Patna, India: NLP Association of India (NLPAI), Dec. 2020, pp. 362–367. [Online]. Available: https://aclanthology.org/2020.icon-main. 49

[16] S. Parida, E. Villatoro-Tello, S. Kumar, P. Motlicek, and Q. Zhan, "Idiap submission to swiss-german language detection shared task." in SwissText/KONVENS, 2020.

[17] Z. Wei, D. Miao, J.-H. Chauchat, R. Zhao, and W. Li, "N-grams based feature selection and text representation for chinese text classification," International Journal of Computational Intelligence Systems, vol. 2, no. 4, pp. 365–374, 2009.

[18] A. Kulmizev, B. Blankers, J. Bjerva, M. Nissim, G. van Noord, B. Plank, and M. Wieling, "The power of character n-grams in native language identification," in Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 2017, pp. 382–389.

[19] K. Watanabe and M. Goto, "Lyrics information processing: Analysis, generation, and applications," in Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA), 2020, pp. 6–12.

[20] E. Jocelin and T. Tryana, “Code mixing and code switching in a korean-song lyric,” Lexeme: Journal of Linguistics and Applied Linguistics, vol. 1, no. 2, 2019.
[21] K. Liew, Y. Uchida, N. Maeura, and E. Aramaki, “Classification of nostalgic music through lda topic modeling and sentiment analysis of youtube comments in japanese songs,” in Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA), 2020, pp. 78– 82.