Automatic Dialect Detection for Low Resource Santali Language

Sunil Kumar Sahoo
GIET University
Gunupur, Odisha, India
sunil.sahoo@giet.edu

Brojo Kishore Mishra
GIET University
Gunupur, Odisha, India
bkmishra@giet.edu

Shantipriya Parida
Idiap Research Institute
Martigny, Switzerland
shantipriya.parida@idiap.ch

Satya Ranjan Dash
KIIT Deemed to be University
Bhubaneswar, Odisha, India
sdashfca@kiit.ac.in

Jatindra Nath Besra
North Orissa University
Baripada, Odisha, India
jatindra.nathbesra@gmail.com

Esaú Villatoro Tello
Universidad Autónoma Metropolitana
Mexico City, Mexico
evillatoro@cua.uam.mx

***Abstract*—Among different natural language processing (NLP) tasks, language detection is considered a difficult task especially for similar languages, varieties, and dialects. To build any tools or technologies in NLP, it is required reliable and robust language detection tool. Although such language detection tools and technologies are available for many high-resource languages, it is difficult to find them for low-resource languages due to the lack of training datasets. In this work, we compiled a language detection dataset for the low resource language *Santali*. We considered the *Sanatali* written in *Odia* script and standard *Odia* script. We used a deep supervised autoencoder for language detection. The deep supervised autoencoder can detect both *Odia* and *Santali* languages although written using the same *Odia* script. We obtained an accuracy of 100% on the test set. Our proposed dataset will be enriched *Santali* language and help researchers to build tools and technologies for this low resource language.**

***Index Terms*—Natural Language Processing, Language Detection, Dialect Detection, Supervised Autoencoder.**

I. Introduction

With the growth of the World Wide Web (WWW), the data generated by web users are growing rapidly. Day by day, more and more people are accessing the web and more languages and dialects are gaining representatives within WWW which need to be processed. Thus identification of language is the key component in building various natural language processing (NLP) resources [1]. Language detection is the task of determining the language for the given text. Although it has progressed substantially still few challenges exist, namely: (1) detection of similar languages; (2) detection of languages when multiple language contents exist in a single document; and (3) short text detection [2]–[4].

It is a difficult task to discriminate between very close languages or dialects, for example, German dialect identification, Indo-Aryan language identification [5]. Although dialect identification is commonly based on the distributions of letters or letter n-grams, it may be unable to distinguish related dialects with very similar phoneme and grapheme inventories for some languages [6].

India is a multi-lingual country with great linguistic and cultural diversities, starting from Jammu and Kashmir in the north to Kanyakumari in the south and Gujarat in the west to Arunachal Pradesh in the east, people speak different languages and dialects. A few of them are Hindi, Bengali, Odia, Tamil, Telugu, Kannada, and Urdu, etc. Indian constitution considers English and Hindi as the official languages. In 2015, the Government of India launched a massive campaign named "Digital India". This was done to make the government services accessible in various parts of the country. The main aim was to improve access to technology for the people of the country. But most of the people in India they cannot read, write and speak Hindi and English. The state of indigenous languages today mirrors the situation of indigenous peoples. In many parts of the world, they are on the verge of disappearance. Preserving and building NLP resources for the indigenous languages are in the focus now and many researchers are working in this direction [7]. The biggest factor contributing to their loss is state policy. But today, the major influence is that their languages are threatened. Santali language is one of the indigenous languages. Santhals are the largest Adivasi (indigenous) community in the Indian subcontinent with a population of more than 10 million, and they reside mostly in the Indian states of Jharkhand, Orissa, West Bengal, Assam, and Bihar, and sparsely in Bangladesh and Nepal. Santali was an oral language until 1925, until the development of *Ol Chiki* script. Apart from *Ol Chiki* script, Santali is written in *Devanagari*, *Bengali-Assamese*, *Odia*, and *Roman script*. Although the Santali language has a rich cultural heritage, the lack of digitization, makes it inaccessible for further development of technology, resulting in a complete lack of online presence. Consequently, NLP tools and technologies are urgently needed for this language.

Accordingly, the main objectives of this paper are:

1) To create a dataset for Santali dialect detection. Particularly Santali written in Odia script.
2) Evaluate the performance of deep machine learning techniques for dialect detection.

II. Related Work

Studying dialects started early in 1877, when George Wenker was conducting a series of surveys to identify dialect regions [8], [9]. Then Bailey [1] carried out one of the first attempts to define the Midland dialect and whether it exists or not. This study concluded that identifying dialects should not depend on vocabulary, as this may vary according to community or class within the same geographic region. Following the same route, Davis and Houck [10] were also trying to find out whether the Midland region can be treated as a separate region of dialect or not. Their study was successful in extracting the phonological and lexical features among 11 cities that lie on a north-south line [10]. The conclusion was that the Midland region cannot be considered to be a transition region and that there is a linear relationship between the distant South and Southern dialects [9]. In opposition to the latter hypothesis, Johnson showed that combining phonological and lexical features was wrong, as doing so affects the data patterns negatively and thus yields incorrect results [11]. Using some words, the proof was produced that there is a clear difference between both the North and Midland dialects and the South and Midland dialects, but not between the dialects of the North and the South.

Similarly, German Dialect Identification was another subtask for the competition. Works on German Dialect Identification are lesser in number when compared to Arabic. One of the initial works uses the n-gram method at a character level for this [12]. Another very different approach included recognizing the high-frequency words in the data [13]. The dataset that we use here consists of speech transcripts provided by the VarDial 2017 subtask taken from ArchiMob Corpus of Spoken Swiss German released in 2016 [14]. The best results in this task were shown by a meta-classifier based on SVM classifiers [15].

Santali is written in *Ol Chiki* script and spoken by 7.6 million people, according to the 2011 census [2], in India alone and its speakers live mostly in Jharkhand, West Bengal, Odisha, and Assam. It is also spoken in Bangladesh and Nepal. According to the 2011 Census, carried out by Registrar General and Census Commissioner, India, the total tribal population (Schedule Tribes) of India stands at 8.6 percent in 27 out of 29 states and 3 out of 7 Union Territories. Santhals are the third largest and most advanced tribe in India. The literacy rate of Santali approximately rests at 53.11% (Census 2011). In 2004, Santali was included in the 8th schedule of the Indian Constitution as a scheduled language. Their dialect belongs to the Munda-Austro-Asiatic group and is derived from the old Kherwali language, spoken mainly in Odisha, West Bengal, Jharkhand, Assam, Bihar, Tripura, and Mizoram. The aboriginal Santali's developed *Ol Chiki* as their script in the first half of the 20th century approximately in 1925. The script was developed by the dedicated Santali linguist Raghunath Murmu. *Ol Chiki* has 30 alphabets and is written from left to right. The script is made up of 6 vowels and 24 consonants, along with 5 basic diacritics. The shapes of the alphabets are inspired by nature, life, and physical forms that are present in the Santali habitat. There is a unique set of digits that represent numbers that also uses the decimal number system. The punctuation symbols used are mostly borrowed from the Latin Roman language. The dialect of the Santali language is the most complex aspect of speech recognition, as it pertains to the language characteristics of a specific regional community. Research in the field of dialect linguistics is still limited due to the unavailability of databases and the time-consuming analysis process. As technology has advanced a robust speech recognizer - one which can handle unstable conditions, such as noise and accent variation - has become an urgent need. Dialects are not static; they change with place and time. In other words, as dialects change with geographic boundaries they may also vary due to language changes over time. Today, we see that new generations use phrases and vocabulary that were not in existence or use in the past. Therefore, languages change over time and their dialects change as well. Also, anyone language includes several dialects. For example, India has many dialects of the Santali language. Different dialects exist in the different states of the Santali language in India.

[1] americanspeech.dukejournals.org/content/78/3/307.refs

[2] https://www.censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf

III. Proposed Method

The overall architecture of the proposed method is shown in Figure 1. The following subsections briefly describe the main components of our approach. We used supervised autoencoder which found very effective for the language and dialect detection recently [16], [17]. We adapted the model used by [16].

A. Supervised Autoencoder

A supervised autoencoder (SAE) is an autoencoder (AE) with the addition of a supervised loss on the representation layer. In the case of a single hidden layer, a supervised loss is added to the output layer and for a deeper AE, the innermost layer has a supervised loss added to the bottleneck layer that which usually transferred to the supervised layer after training the AE.

In supervised learning, the main goal is to learn a function for a vector of inputs $\vec{x} \in \mathbb{R}^d$ to predict a vector of targets $\vec{y} \in \mathbb{R}^m$. Consider SAE with a single hidden layer of size k, and the weights for the first layer are $\vec{F} \in \mathbb{R}^{k \times d}$. The function is trained on a finite batch of independent and identically distributed (i.i.d.) data, $(\vec{x}_1, \vec{y}_1), ..., (\vec{x}_t, \vec{y}_t)$, with the goal of a more accurate prediction on new samples generated from the same distribution. The weight for the output layer consists of weights $\vec{W}_p \in \mathbb{R}^{m \times k}$ to predict $\vec{y}$ and $\vec{W}_r \in \mathbb{R}^{d \times k}$ to reconstruct $\vec{x}$. Let L_p be the supervised loss and L_r be the loss for the reconstruction error. In case of regression, both losses might be represented by a squared error, resulting in the objective:

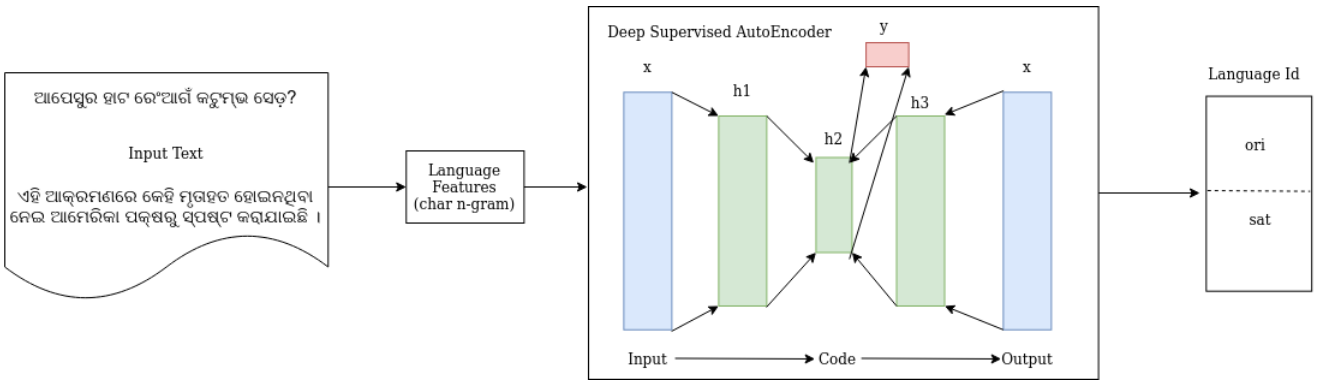

Fig. 1: Proposed model architecture. The extracted features of the text are input to the supervised autoencoder. The target "y" are included. The classification output are the language id for the classified languages (ori: Odia, sat: Santali).

$$\frac{1}{t}\sum_{i=1}^{t}\left[L_p(\vec{W}_p\vec{F}\vec{x}_i,\vec{y}_i)+L_r(\vec{W}_r\vec{F}\vec{x}_i,\vec{x}_i)\right]= \frac{1}{2t}\sum_{i=1}^{t}\left[||\vec{W}_p\vec{F}\vec{x}_i-\vec{y}_i||_2^2+||\vec{W}_r\vec{F}\vec{x}_i-\vec{x}_i||_2^2\right] \quad (1)$$

The addition of supervised loss to the AE loss function acts as a regularizer and results (as shown in equation 1) in the learning of the better representation for the desired task [18]. In summary, an SAE represents a neural network that jointly predicts targets and inputs.

B. Bayesian Optimizer

In the case of SAE, there are many hyperparameters related to model construction and optimization. AE training and performance often benefit from hyperparameter tuning to avoid over and under-fitting.

Bayesian optimization (BO) is a state-of-the-art hyperparameter optimization algorithm that reached competitive performances on several optimizations benchmarks [19], [20]. BO is a technique based on Bayes theorem to direct a search for a global optimization problem that is efficient and effective. It works by building a probabilistic model of the objective function, called the surrogate function, that is then searched efficiently with an acquisition function before candidate samples are chosen for evaluation on the real objective function.

C. Textual Features

Character n-grams are fed as an 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 additionally capable of detecting the morphological makeup of a word [21], [22]. The extracted n-gram features are input to the deep SAE as shown in the Figure 1. The deep SAE contains multiple hidden layers. Hyperparameters were optimized using BO.

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

TABLE I: Search space hyper parameter range.

IV. Experimental Setup and Datasets

A. Hyperparameters

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

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

TABLE II: SAE model configurations for the dataset.

B. Datasets

Due to the unavailability dataset, we created the labeled dataset by selecting Odia and Santali written in the Odia script. The sample of dataset text is shown in the Figure 2. We selected 500 Odia sentences from *Samanantar*, the largest parallel corpora collection for 11 Indian languages [23]. The parallel corpora collection includes English-Odia parallel text that covers many domains. We selected the Odia sentences of word length between 5 to 15 words per sentence. We extracted 478 Santali sentences from online and books. The source of the sentences selected for the Santali is shown in Table III. We used OCR for extracting contents from the book with manual verification and correction.

After merging both datasets with labeled 'sat' for Santali and 'ori' for Odia based on the corresponding ISO language

Source	Type
Santali-English book (A. Campbel)	Book
Odisha Virtual Academy	Web (online portal)
A Concise English-Santali Book (R. C. Hansdah)	Book
Sahaja Santhali Sikshya By Dr. Damayanti Besra	Book

TABLE III: Santali Data Source.

Dataset	Training	Development	Test
Odia-Santali	782	98	98

TABLE IV: Dataset Statistics.

code [3]. For experiment, we divided into 80:10:10 ratio for the train/dev/test set as shown in Table IV. During preprocessing we have removed the full-stop mark in Odia i.e. Odia danda mark (।) from the Odia text and (.) symbol from the Santali text.

Language	Santali/Odia Text with Transliteration	English Translation
Santali	ᱟᱢ ᱚᱞᱚᱜ ᱯᱟᱡᱦᱟᱣ ᱮᱢ ᱵᱚᱫᱚᱭ ᱠᱮ? am olog pazhaw em badaya ce?	Do you know how to read and write ?
	ᱥᱟᱱᱚᱢᱟᱜ ᱱᱚᱯᱟᱭ ᱜᱮᱛᱚ? sanamag napay geTo?	Are you fine ?
Odia	ଲୋକସଭା ନିର୍ବାଚନ ପୂର୍ବରୁ ପ୍ରଧାନମନ୍ତ୍ରୀ ମୋଦିଙ୍କ ପ୍ରୟାଗରାଜ ଗସ୍ତକୁ ମହତ୍ତ୍ୱପୂର୍ଣ୍ଣ ବୋଲି କୁହାଯାଉଛି । lōkasabhā nirbācana pūrbaru pradhānamantrī mōdiṅka prayāgarāja gastaku mahatt᷾pūrṇṇa bōli kuhāyāuchi.	The visit of Prime Minister Modi is obviously going to be crucial as the visit comes ahead of the elections in the state.
	ଖବର ପାଇ ପୁଲିସ ଘଟଣାସ୍ଥଳରେ ପହଞ୍ଚି ମୃତଦେହ ଜବତ କରି ଘଟଣାର ତଦନ୍ତ ଚଲାଇଛି । khabara pāi pulisa ghaṭaṇāsthaḷarē pahañci mṛtadēha jabata kari ghaṭaṇāra tadanta caḷāichi.	Police reached the spot and started investigation after seizing the dead body.

Fig. 2: Sample Santali and Odia text.

The statistics of the dataset are shown in Table IV.

C. Baseline

As a baseline system, we implemented a binary linear Support 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.

V. Results and Discussion

The SAE and SVC model's performance in terms of classification accuracy is shown in Table V.

		Accuracy	
Model	Dataset	Dev	Test
SAE (char-3gram)	Odia-Santali	100%	100%
SVC (tf-idf)	Odia-Santali	97%	97%

TABLE V: Overall performance of the proposed approach.

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

The confusion matrix on the dev and test are shown in Figure 3 and Figure 4 respectively.

Based on the results, it can be seen that the model performs well on both the dev set (100%) and the test set (100%). Few sample output from the test set is shown in Figure 5.

[3]https://www.loc.gov/standards/iso639-2/php/code_list.php

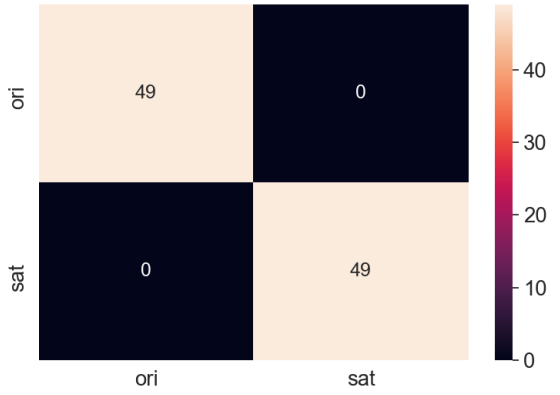

Fig. 3: Confusion matrix for dev set. The language code 'ori' represents for Odia and 'sat' for Santali.

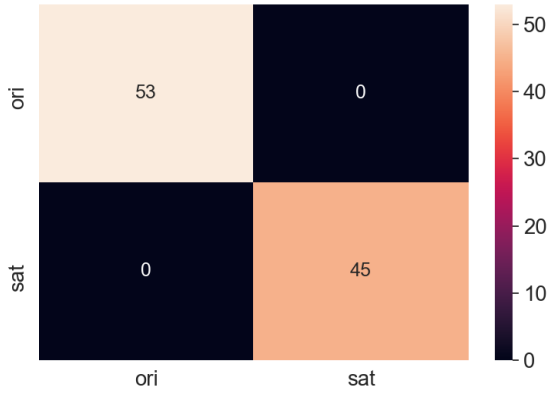

Fig. 4: Confusion matrix for test set. The language code 'ori' represents for Odia and 'sat' for Santali.

Test set	Prediction	Remark
[Santali Text] ᱟᱯᱮᱥᱩᱨ ᱦᱟᱴ ᱨᱮᱱᱟᱜ ᱡᱩᱴᱩᱢ ᱪᱮᱰᱩ? [Transliteration] ape sur hat renag quTum ceD? [English Translation] What is the name of the nearest market ?	Santali	Correct
[Santali Text] ᱮᱞᱟᱜ ᱡᱟᱱᱚᱣᱟ ᱟᱨ ᱪᱮᱢᱮ ᱳᱡᱟᱜ᱾ᱨᱮ ᱪᱮᱫᱟᱜ ᱯᱮ ᱫᱳᱦᱳ ᱠᱚᱣᱟ? [Transliteration] etag janova ar ceMe ozagre ceDag pe Doho kowa? [English Translation] Why do you keep other birds and animlas in your house ?	Santali	Correct
[Odia Text] ଏହି ଆକ୍ରମଣରେ କେହି ମୃତାହତ ହୋଇନଥିବା ନେଇ ଆମେରିକା ପକ୍ଷରୁ ସ୍ପଷ୍ଟ କରାଯାଇଛି । [Transliteration] ēhi ākramaṇarē kēhi mṛtāhata hōinathibā nēi āmērikā pakṣaru spaṣṭa karāyāichi . [English Translation] The United States has said no one was killed in the attack	Odia	Correct

Fig. 5: Sample test set result.

	Precision	Recall	F1
Odia (ori)	1.00	1.00	1.00
Santali (sat)	1.00	1.00	1.00

TABLE VI: Model (SAE) performance for each class (Precision, Recall, and F1) on test set.

VI. Conclusion

In this paper, we propose an approach for language detection of the Santali language. Also, we provided a dataset consisting of Odia and Santali suitable for language detection research and building NLP tools and technologies. The Odia Santali dialect detection dataset is available at:

https://github.com/shantipriyap/
Odia-Santali-Dialect-Detection-Dataset

The future work includes *i)* Extend the dataset with more Santali text, *ii)* experiment with other similar dialects of Odia for performance evaluation of the deep supervise autoencoder.

References

[1] T. Kocmi and O. Bojar, "Lanidenn: Multilingual language identification on character window," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 927–936.

[2] I. Balazevic, M. Braun, and K.-R. Müller, "Language detection for short text messages in social media," *arXiv preprint arXiv:1608.08515*, 2016.

[3] M. Lui, J. H. Lau, and T. Baldwin, "Automatic detection and language identification of multilingual documents," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 27–40, 2014.

[4] J. Williams and C. Dagli, "Twitter language identification of similar languages and dialects without ground truth," in *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 2017, pp. 73–83.

[5] T. Jauhiainen, K. Lindén, and H. Jauhiainen, "Language model adaptation for language and dialect identification of text," *Natural Language Engineering*, vol. 25, no. 5, pp. 561–583, 2019.

[6] Y. Scherrer and O. Rambow, "Natural language processing for the swiss german dialect area," in *Semantic Approaches in Natural Language Processing-Proceedings of the Conference on Natural Language Processing 2010 (KONVENS)*. Universaar, 2010, pp. 93–102.

[7] S. Parida, S. Panda, A. Dash, E. Villatoro-Tello, A. S. Doğruöz, R. M. Ortega-Mendoza, A. Hernández, Y. Sharma, and P. Motlicek, "Open machine translation for low resource South American languages (AmericasNLP 2021 shared task contribution)," in *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Online: Association for Computational Linguistics, Jun. 2021, pp. 218–223. [Online]. Available: https://aclanthology.org/2021.americasnlp-1.24

[8] J. K. Chambers and P. Trudgill, *Dialectology*. Cambridge University Press, 1998.

[9] A. A. Nti, "Studying dialects to understand human language," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[10] L. M. Davis and C. L. Houck, "Is there a midland dialect area?—again," *American speech*, pp. 61–70, 1992.

[11] E. Johnson, "Yet again: The midland dialect," *American speech*, vol. 69, no. 4, pp. 419–430, 1994.

[12] W. B. Cavnar, J. M. Trenkle *et al.*, "N-gram-based text categorization," in *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, vol. 161175. Citeseer, 1994.

[13] N. Ingle, *A language identification table*. Technical translation international, 1980.

[14] T. Samardzic, Y. Scherrer, and E. Glaser, "Archimob-a corpus of spoken swiss german," in *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. European Language Resources Association (ELRA), 2016.

[15] M. Zampieri, S. Malmasi, N. Ljubešić, P. Nakov, A. Ali, J. Tiedemann, Y. Scherrer, and N. Aepli, "Findings of the vardial evaluation campaign 2017," in *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects*, 2017.

[16] S. Parida, E. Villatoro-Tello, S. Kumar, M. Fabien, and P. Motlicek, "Detection of similar languages and dialects using deep supervised autoencoders," in *Proceedings of the 17th International Conference on Natural Language Processing*, no. CONF. ACL, 2020.

[17] S. Parida, E. VILLATORO-TELLO, S. Kumar, M. Fabien, and P. Motlicek, "Detection of similar languages and dialects using deep supervised autoencoders," in *Proceedings of the 17th International Conference on Natural Language Processing*, 2020.

[18] L. Le, A. Patterson, and M. White, "Supervised autoencoders: Improving generalization performance with unsupervised regularizers," in *Advances in Neural Information Processing Systems*, 2018, pp. 107–117.

[19] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

[20] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of machine learning research*, vol. 13, no. Feb, pp. 281–305, 2012.

[21] 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.

[22] 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.

[23] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, R. AK, A. Sharma, S. Sahoo, H. Diddee, D. Kakwani, N. Kumar *et al.*, "Samanantar: The largest publicly available parallel corpora collection for 11 indic languages," *arXiv preprint arXiv:2104.05596*, 2021.