GRINREY

Advanced Research in Computer Engineering

Sandip A. Kale
Editor

Research Transcripts in Computer, Electrical and Electronics Engineering

2 Elucidation of Bangla Language Classification Using Neural Network Approach

Pritom Sarker[a, *] and Jannatul Ferdous[a], Nakib Aman Turzo[a]

[a]Computer Science & Engineering Department, Varendra University, Rajshahi, Bangladesh

*Corresponding author: me.pritom@gmail.com

ABSTRACT

Bangladesh has two principal languages called Sadu and Cholit. In the early times, Sadhu was operational and was composed of Sanskrit components but the current era has shifted to Cholit language, which is now being used most commonly. Sadhu was mostly used for formal documentation purposes and it is the need of the hour to translate them to Cholit language because it is more speaker friendly and can be easily understandable. Therefore, in this chapter efforts were done to familiarize the current era with the Sadhu language by creating software. Few sentences were selected and the final dataset was obtained by Principal Component Analysis (PCA). Python is used for different machine learning algorithms. Maximum work was done on Scikit Learn which is Term Frequency-Inverse Data Frequency (TF-DF) Vectorizer's class. The best performance was given by Neural Network with high precision. Speed was also anticipated and values were determined through graphs. The results showed that it translated all words from Sadhu to Cholit efficiently and in a well-oriented way. Therefore, Sadhu's complexity has been removed in this era.

Keywords: Machine learning, Neural network, Principal Component Analysis, Python, Term Frequency

Advanced Research in Computer Engineering | ISBN: 978-81-948951-2-1
Book Series: Research Transcripts in Computer, Electrical and Electronics Engineering | Vol: 02

1. INTRODUCTION

Sadhu language is an elaborated bygone register related to Bengali dialect, which was used most commonly used in the course of the Bengali Renaissance from the 19th to 20th century. It differs in vocabulary structure; verb form and its main components are Sanskrit and Tasama. Unlike Cholit language, Sadhu was utilized mostly for penmanship, which is indigenous but used in longhand and also in verbalized form. Diglossia is the main section under which the two languages lie. Most writings now come under the Cholit language. In some parts of Bangladesh like Chittagong is superficially identical to Cholit Bangla. In ancient times, the Sadhu language was used informal documentation but now it is outdated in the current situation. The origin of the Sadhu language in the literature was because of the intellectuals of Gour, which first used it. Due to this reason, it is called the Sadhu Gouriyo language. Cholit language used presently is the most speaker friendly language and is the most common one. It is the most common bond for communication and understanding for Bengali speakers. In this chapter, our focus is to distinguish between Bengali sentences whether it is Sadhu and Cholit. This effort will open doors for the conversion of Sadhu to Cholit and can translate both languages for the ease of any speaker. It may lead to the creation of software, which can detect whether a sentence is in Sadhu or Cholit and can translate it. The chief goal is to increase the recognition of our generation to ancient literature through these languages by interconversion.

2. LITERATURE REVIEW

Differences in grammar can be revealed by classifiers but cognition can't be determined by it. Five sortal classifiers are utilized by Cantonese than Mandai. 40% of nouns appear without classifier and sortal was taken by 18% of Cantonese and by 3% of Manian [1].

Components of an NP consist of just a classifier using semantic criteria to override their syntactic distributor in Mandarian and Cantonese [2].

A major part of Natural Language Processing is machine translation for the translation of language. Translation consists of three parts which are translation models, language models, and a decoder. For the translation of English to Hindi, a statistical machine translation system was developed. In a Linux environment, the model is developed by using this software [3].

Categorization of speech and language processing systems can be done for predefined linguistic information use and it is driven by data. It causes the use of methods of machine learning for extraction automatically and processes corresponding units of information which are indexed as appropriate. Hence,

ALISP (Automatic Language-Independent Speech Processing) was used to exploit an idea with a focus on speech processing [4].

In an ignition, it was depicted that context-free grammar was the main problem in many speech understanding systems. Computationally augmented phrase structure grammar is very demanding. Most efficient finite set grammar can't represent the relation of sentence meaning. It was elaborated by developing an APSG coupled with language analysis and then derived automatically. An efficient translation system was developed using it which was fast compared to others [5].

A paper discussed the integration of natural language and speech processing in Phi DM-Dialog and also kept into consideration a cost-based scheme of ambiguity resolution. With the help of an incremental pairing and generation algorithm, the simultaneous interpretation capability was made [6].

Conversion of language is the most difficult task and a case study was done for this purpose. This involves the conversion of clients system in proprietary language to programming language. There are many factors that were taken into account and affect language conversion at the automation level [7].

CJK Dictionary Publishing Society has launched an investigative project in 1996 for the development of sophisticated segmentation by collaborating with basic techniques for making an elaborated simplified Chinese and traditional Chinese database with 100% precision [8].

Speech to text conversion was done in few pieces of researches for people with impairments in hearing. Software was created which aids using phonetics to correct pronunciations. This helps in the potential of recognition in English hearing [9].

A generic method was introduced to translate to discretized Modern Standard Arabic from a written Egyptian colloquial sentence. It can easily be applied to other dialects of Arabic. A lexical acquisition was done for colloquial Arabic for the conversion of written Egyptian Arabic to MSA [10].

For recognition of two speakers in each of Spanish and English, a system was also developed. It was limited to 400 words. Analysis of language and speech recognition are coupled strongly by using the same language model [11].

In an effort by utilizing neural network conversion of text which was in Hindi to speech was done. It has many applications in the case of blind people and can also be used for educating students. A document containing Hindi text was used as input and a neural network was used for recognizing characters [12].

Grammatical errors have restrictions in variability and function in the historical periods of the English language. They become more productive in the 19th and 20th century along with major extensions in functions, range of lexical associations, and variants [13].

For conversion of Hindi text to speech in Java Swings, a Graphical User Interface was developed. Because it consists of a variety of languages spoken in different areas [14].

Progress was made recently in the synthesis of speech by developing speech synthesizers that have very high intelligibility but there are some problems like lack of naturalness and sound quality. But quality has reached a suitable level in many applications [15].

With embedded spelled letters there are also investigations done for recognition accuracy of speech. There are various methods proposed for localization of spelled letter segments, and to again classify them with a specialized letter recognizer [16].

For the translator software which offsets the absence of educational tools to some extent a development report was prepared because this tool is the need of hearing-impaired person for communication. For enhancing written language skills this tool can be utilized [17].

By using lexicon, rule-based and, data-driven techniques words can be converted to triplets with the aid of a software system which causes the conversion between graphemes and phonemes. In a hybrid system, a shotgun integrates these techniques and adds educational and linguistic information regarding graphemes and phonemes [18].

A speech-to-text engine was developed online in real time for the transfer of speech into written language and also it requires special techniques [19].

Translation dilemmas examination was done in a qualitative research. By the considering implications of the same problem's medium of spoken and written languages was critically challenged. The translation was centered and the way of dealing with issues raised by a representation would be the main concern for all researchers [20].

Sadhu and Cholit languages were also classified using machine learning techniques in which Linear Discriminant Analysis (LDA) had given the most precise and accurate results and has opened doors for further research by using different approaches [21].

Language identification of similar languages was done using Recurrent Neural Networks and its improved accuracy in recognizing different languages by combining word vector representation and Long Short-Term Memory [22].

No significant study has been found on the classification of Sadhu and Cholit language sentences and conversion. As a result, this study is conducted based upon similar works conducted in other languages.

3. METHODOLOGY

For amassing sentences related to Sadhu and Cholit data was gathered from ancient Literature books. From five significant literature books 2483 Sadhu sentences and from six important literature books 3508 Cholit sentences were collected for this purpose. The detailed procedure is depicted in Fig 1.

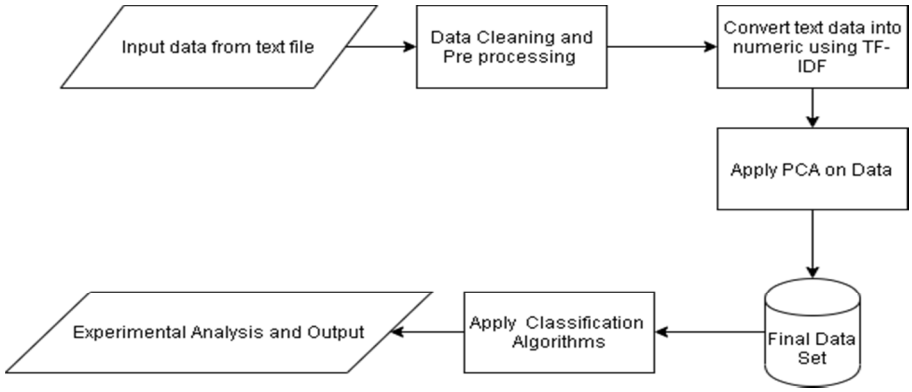


Fig. 1. Work Flow of the classification process

1. Data Clean

Non- English words are (before or after processing of natural language data it got filtered) in the data set. Extraction of all Non-English words was performed first. The natural language toolkit information center of Python is being used in this regard. The information set contains all non-English sentences. After going through the whole process 1983 data set was derived. For numeric categorization, Sadhu is termed as numeric 0 and Cholit as numeric 1.

2. Short Term Frequency–Inverse Data Frequency

An analytical statistic is a numerical or scientific form of statistic which is being contemplated to mirror the principle of a word in a docket or corpus and is called Short Term Frequency-Inverse Data Frequency (TF-IDF). This factor has weightage in retrieving information, text mining, and user modeling through the hunting of this data.

3. **Term Frequency (TF)**

Frequency of a word that pops up in a docket divided by the gross number of words in the document. Every document has its term frequency.

4. Inverse Data Frequency (IDF)

The log of the document numbers is divided by word w containing documents. Inverse data frequency determines the weight of rare words across all documents in the corpus.

Most work is being done from Scikit-Learn which is TF-DF Vectorizer's class. Text data was taken by it and converted to a numeric information set. After this conversion, the data has 3394 features. Features extraction was performed using PCA.

5. Principal Component Analysis

Through orthogonal linear transformation, data is converted to a new coordinate system so in an ordered manner each coordinate has the greatest variance by scalar projection

and so on hence, known as principal component analysis. The Principal component analysis is a section that comes under the class of Scikit-learn. First coordinate is known as the first principal component with higher variance in it and lower variance comes under the second coordinate. After application of principal component analysis, our information has 1678 traits. Data quality got lost when applications of dimensions of principal component analysis got reduced.

95% caliber of data was being maintained in the principal quality analysis case. By setting the value of 'n' components as 0.95 the amount of quality of real data preserved is 95%. 1678 characters remained in our data after the application of principal component analysis.

'তিনি সাহিত্যিক সুতরং টাকার কথা তুলিয়া তাঁহার সারস্বত সাধনার অমর্যাদা করিতে চাই না	Sadhu
'[হায় রে পোড়াকপালে সাহিত্যিক!] কিন্তু যদি তিনি নাম ধাম বদল করিয়া এই হাঁরা হরণের গল্পটা লিখিতে পরেন তাহা হইলে ত	Sadhu
'শ্রদ্ধা ও নমস্কার গ্রহণ করিবেন	Sadhu
'ওভারহেড প্রোজেক্টরের সুইসটি অফ করে তিনি হলভর্তি দর্শকদের দিকে তাকালেন	Cholit
'বিজ্ঞানীদের কনফারেন্সে বক্তব্য শেষ হবার পর সাধারণত ছোট একটি সৌজন্যমূলক করতালিদেয়া হয় কিন্তু এবারে একটি বি	Cholit
'এই সেশনটির সভাপতি সেন্টজন বিশ্ববিদ্যালয়ের বৃদ্ধ অধ্যাপক বব রিকার্ডো প্র মে করতালি দিতে শুরু করলেন এবং গ্যালারি	Cholit
'দেখতেদেখতে করতালির প্রচণ্ড শব্দে হলঘরটি ফেটে যাবার উপ্রম হল কিন্তু তবুও সেটি থেমে যাবারকোনো লক্ষণ দেখা গে	Cholit
'বিজ্ঞানীদেরকনফারেন্সে সাধারণত সাংবাদিকরা থাকেন না কিন্তু জেনেটিক ইঞ্জিনিয়ারদের এই বার্ষিককনফারেন্সে আলবার্তে	Cholit
'ফটো তোলা সম্পূর্ণ নিষিদ্ধহওয়া সত্ত্বেও সাংবাদিকদের ক্যামেরা ফ্ল্যাশ জ্বলতে শুরু করল এই ঐতিহাসিক মুহূর্তটি ধরেরাখার	Cholit
'বৃদ্ধ অধ্যাপক বব রিকার্ডো শেষ পর্যন্ত উঠে দাঁড়ালেন তাকে নির্দিষ্ট সময়ের মাঝে সেশনটিশেষ করার দায়িত্ব দেয়া হয়েছে	Cholit
'যদি এখনই তিনি নিয়ন্ত্রণটুকু হাতে না নিয়ে নেন সেটি সম্ভবহবার কথা নয়	Cholit

Fig. 2. Data Set before TF-IDF and PCA

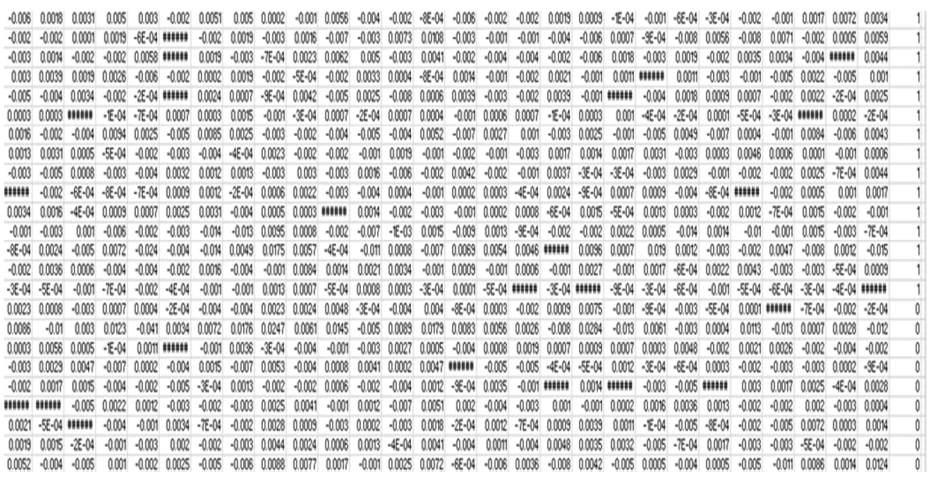


Fig. 3. Processed Data After TF-IDF and PCA

The processed data has 1042 different fields of numeric data in which the last field signifies 1 for the Cholit and 0 for the Sadhu.

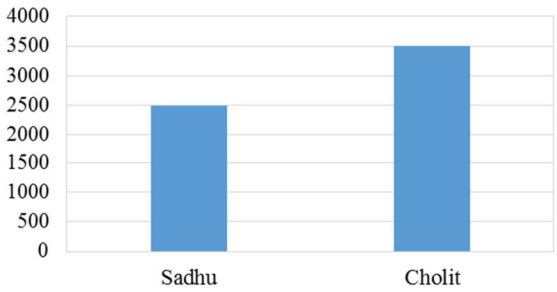


Fig. 4. Number of sentences used

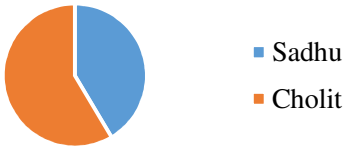


Fig. 5. Relative measurement of the number of sentences used

4. RESULTS AND EXPERIMENTAL ANALYSIS

Fifteen algorithms were used for classification. For this operation, scikit learn and Keras library was used and chosen eight best models best on the cross-validation score by doing it about 10 folds.

Table 1: Accuracy Chart of the top eight performers.

Classifier Name	**Accuracy (%)**
Logistic Regression	73.83
SVM (Linear)	74.04
Ridge Classifier	72.01
Linear discriminant analysis	75.07
AdaBoost	72.11
Standard NN Model	97.76
Smaller NN Model	98.16
Larger NN Model	98.75

Among all classifiers, Neural network functions best as depicted in the graph. In the case of LDA, it is expressed as a dependent variable as a linear combination of other features or measurements and has a resemblance with variance (ANOVA) and regression. As principal component analysis (PCA) and factor analysis, both look for linear combinations of variables which elaborate the data so LDA has also resembled them. When for each observation independent variables are continuous quantities, DA also works there. Discriminant correspondence analysis is an equivalent technique to categorical independent variables.

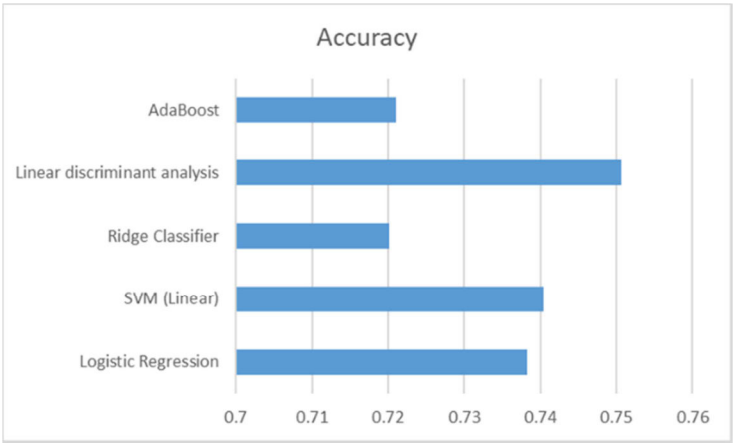


Fig. 6. Accuracy comparison of top 5 performers

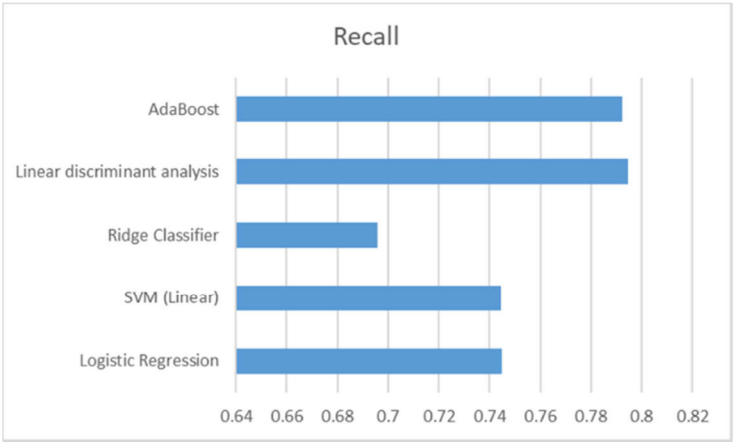


Fig. 7. Recall comparison of top 5 performers

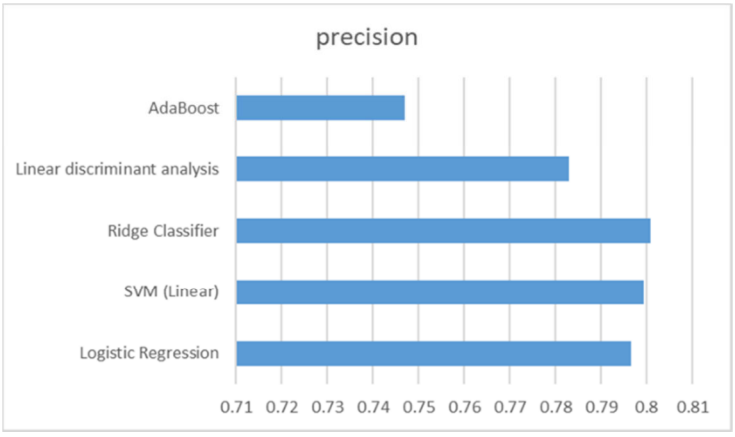


Fig. 8. Precision comparison of top 5 performers

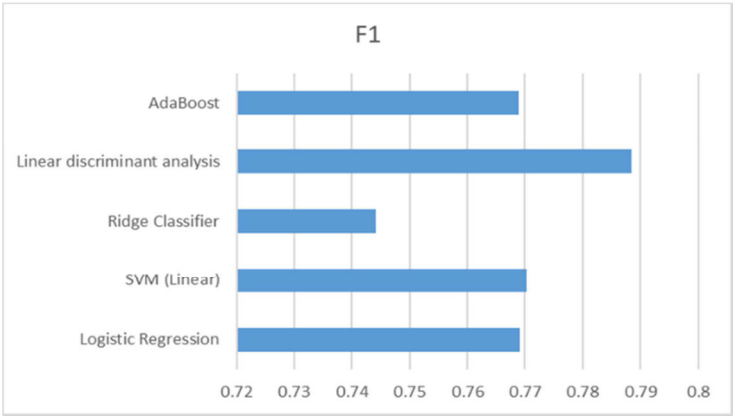


Fig. 9. F1 score comparison of top 5 performers

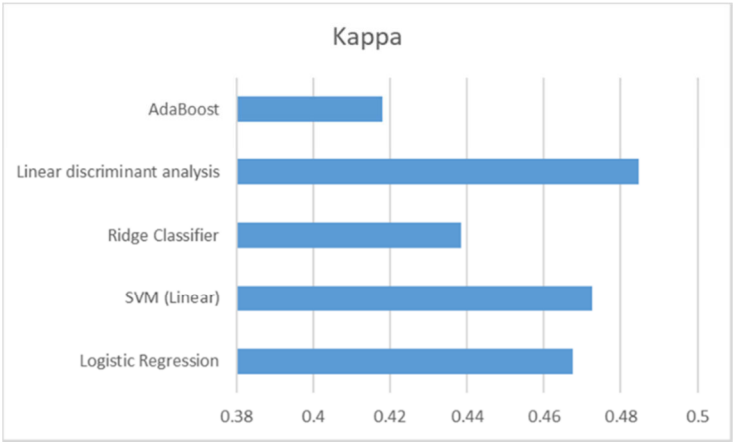


Fig. 10. Kappa comparison of top 5 performers

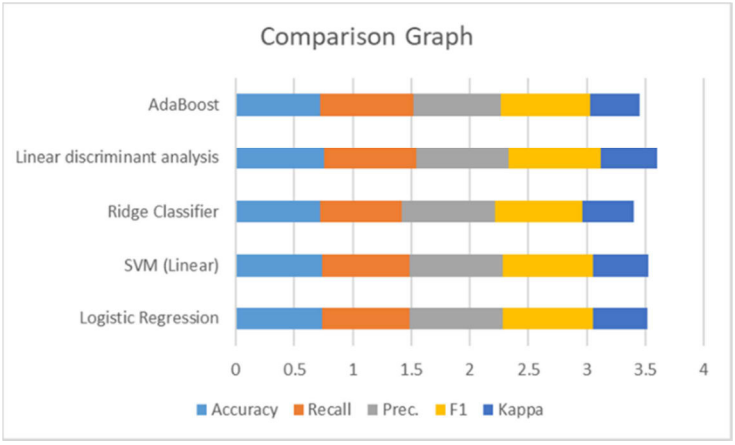
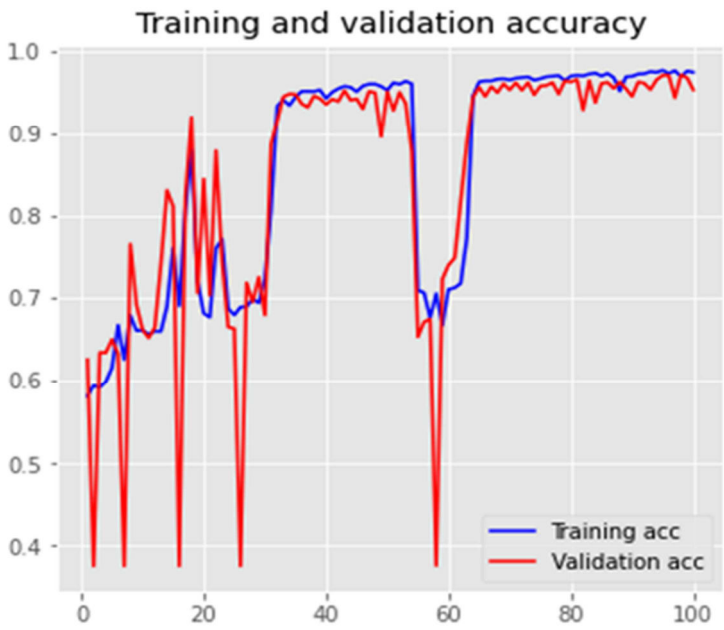
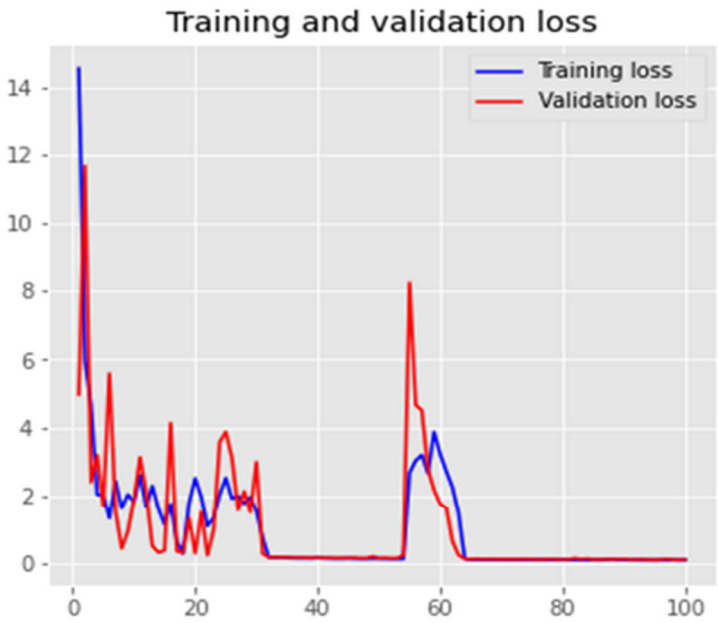


Fig. 11. Overall performance comparison of top 5 performers

4.1. Standard NN Model



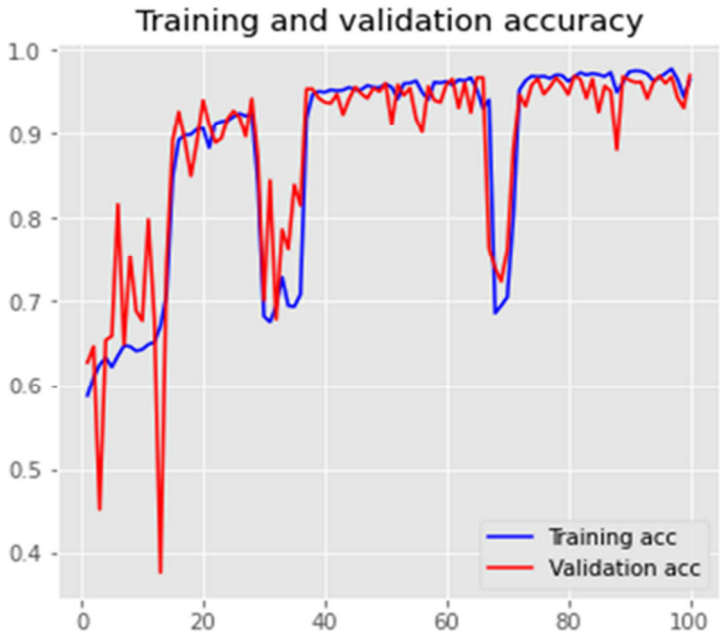
(a) Training Accuracy: 0.9776



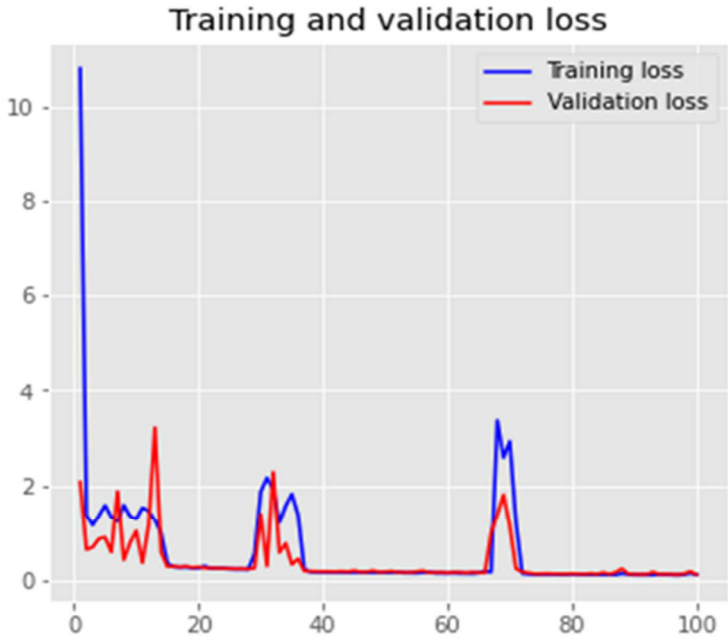
(b) Testing Accuracy: 0.9661

Fig. 12. (a) Accuracy curve using Standard Neural Network (b) Loss curve using Standard Neural Network

4.2. Smaller NN Model



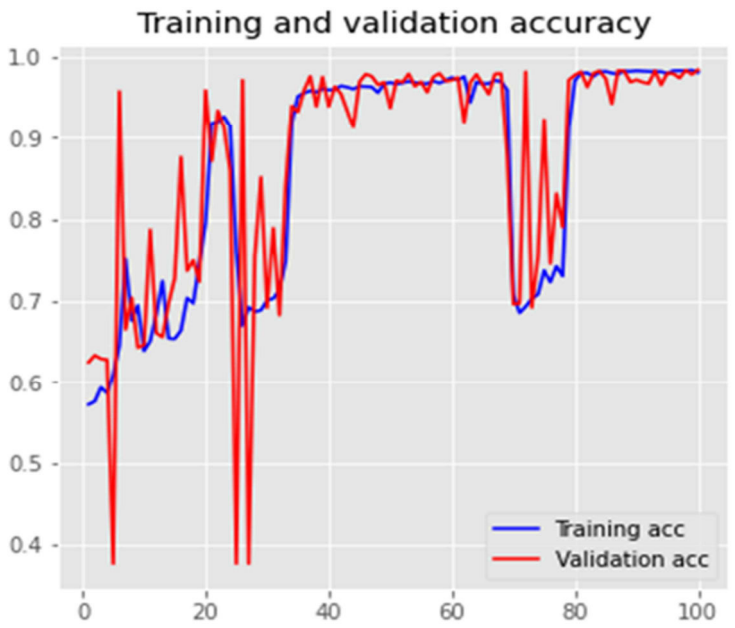
(c) Training Accuracy: 0.9816



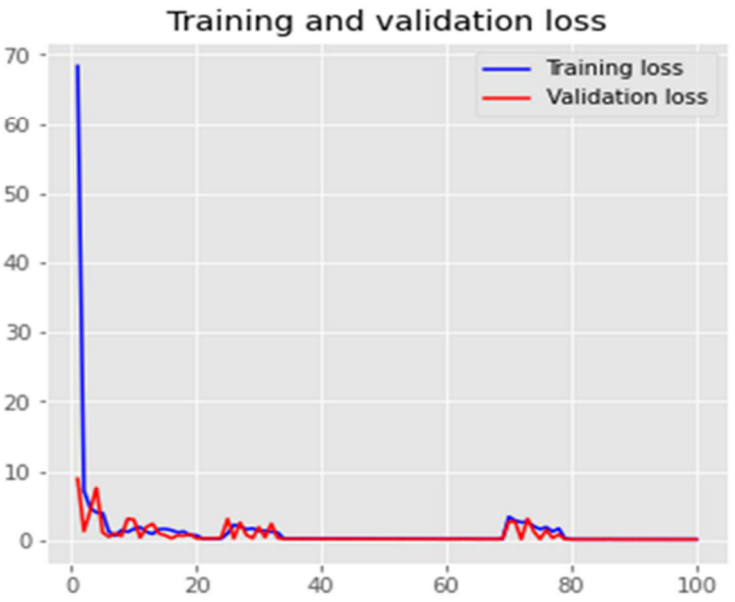
(d) Testing Accuracy: 0.9708

Fig. 13. (a) Accuracy curve using Smaller Neural Network (b) Loss curve using Smaller Neural Network

4.3. Larger NN Model



Training Accuracy: 0.9875



Testing Accuracy: 0.9840

Fig. 14. (a) Accuracy curve using Larger Neural Network (b) Loss curve using Larger Neural Network

5. CONCLUSION

By taking into account the whole process precise results were given by Neural Network in comparison to traditional machine learning approaches and gave 98.75% of accuracy than other models. The classifier can help in cataloging of Sadhu and Cholit languages of Bangladesh. As Bangladeshi literature is enriched with Sadhu language so most of the novels present from the ancient eras are in Sadhu language and this approach can help in their translation more accurately as it's highly uncommon for the present generation. That is why it can be converted to Cholit so that people can get familiar to old literature. RNN based model gave high accuracy than machine learning-based approaches. In future conversion will be tried to convey sadhu into Cholit so that the huge documents and literature can be realized.

REFERENCES

1. M. S. Erbaugh, "Classifiers are for specification: Complementary Functions for Sortal and General Classifiers in Cantonese and Mandarin," *Cahiers de Linguistique Asie Orientale,* vol. 31, no. 1, pp. 36-69, 2002.
2. S. Y. Killingley, Cantonese classifiers: Syntax and semantics, Newcastle upon Tyne: Grevatt & Grevatt, 1983.
3. N. V. p. S. Sharma, "English to Hindi Statistical Machine Translation System," *TIET Digital Repository,* 2 August 2011.
4. G. C. M. Petrovska-Delacrétaz, "Data Driven Approaches to Speech and Language Processing," in *Springer*, Heidelbergh, 2004.
5. D. Roe, F. Pereira, R. Sproat, M. Riley, P. Moreno, and A. Macarron, "Efficient grammar processing for a spoken language translation system," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, CA, USA, USA, 1992.
6. H. Kitano, "Phi DM-Dialog: an experimental speech-to-speech dialog translation system," vol. 24, no. 6, pp. 36-50, 1991.
7. A. Terekhov, "Automating language conversion: a case study (an extended abstract)," in *Proceedings IEEE International Conference on Software Maintenance. ICSM 2001*, Florence, Italy, Italy, 2001.
8. J. a. J. K. Halpern, "Pitfalls and Complexities of Chinese to Chinese Conversion." *international Unicode Conference*, Boston, 1999.
9. M. S. H. Nuzhat Atiqua Nafis, "Speech to Text Conversion in Real-time," *International journal of innovation and scientific research,* vol. 17, pp. 271-277, 2015.
10. H. A. K. S. a. I. Z. Bakr, "A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic."," *international conference on informatics and system,* 2008.
11. RileyAlejandroMacarrón, "A spoken language translator for restricted-domain context-free languages" vol. 11, no. 2-3, pp. 311-319, 1992.

12. P. S. Rathod, "Script to speech conversion for Hindi language by using artificial neural network," in *2011 Nirma University International Conference on Engineering*, Ahmedabad, Gujarat, India, 2011.
13. B. Gray, "Grammatical change in the noun phrase: the influence of written language use," *Cambridge University Press,* vol. 15, no. 2, pp. 223-250, 2011.
14. K. a. R. K. Kamble, "A review: translation of the text to speech conversion for Hindi language."," *International Journal of Science and Research,* vol. 3, 2014.
15. N. a. K. A. Swetha, "Text-to-speech conversion," *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 2, no. 6, pp. 269-278, 2013.
16. A. W. Hermann Hild, "Integrating Spelling Into Spoken Dialogue Recognition," in *European Conference on Speech Communication and Technology*, Germany Carnegie Mellon University, USA, 1995.
17. B. Sarkar, K. Datta, C. D. Datta, D. Sarkar, S. J. Dutta, I. D. Roy, A. Paul, J. U. Molla and A. Paul, "A Translator for Bangla Text to Sign Language," in *2009 Annual IEEE India Conference*, India, 2009.
18. A. N. 1. a. J. Z. Merijn Beeksma 1, "shotgun: converting words into triplets: A hybrid approach to grapheme-phoneme conversion in Dutch," *John Benjamins,* vol. 19, no. 2, pp. 157-188, 2016.
19. P. Khilari, "A Review on Speech to Text Conversion Methods," *Computer Science,* 2015.
20. B. temple, "Qualitative Research and Translation Dilemmas," *Sage Journals,* vol. 4, no. 2, 2004.
21. P. S. B. K. Nakib Aman Turzo, "Interpretation of Sadhu into Cholit Bhasha by Cataloguing and Translation System," *International Journal of Trend in Scientific Research and Development,* vol. 4, no. 3, 2020.
22. M. R. a. M. S. Ermelinda Oro1, "Language Identification of Similar Language using Recurrent Neural Networks," in *10th International Conference on Agents and Artificial Intelligence*, 2018.

Cite this article

Pritom Sarker, Jannatul Ferdous and Nakib Aman, Elucidation of Bangla Language Classification Using Neural Network Approach, In: Sandip A. Kale editor, Advanced Research in Computer Engineering, Pune: Grinrey Publications, 2021, pp. 19-32