International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 3, April 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Interpretation of Sadhu into Cholit Bhasha by Cataloguing and Translation System

Nakib Aman Turzo, Pritom Sarker, Biplob Kumar

Department of Computer Science & Engineering, Varendra University, Rajshahi, Bangladesh

ABSTRACT

Sadhu and Cholit bhasha are two significant Bangladeshi languages. Sadhu was functional in ancient era and had Sanskrit components but in present era cholit took its place. There are many formal and legal paper works present in Sadhu language which direly need to be translated in Cholit because it's more favorable and speaker friendly. Therefore, this paper dealt with this issue by familiarizing the current era with Sadhu by creating a software. Different sentences were chosen and final data set was obtained by Principal Component Analysis (PCA). MATLAB and Python are used for different machine learning algorithms. Most work is being done using Scikit-Learn and MATLAB machine learning toolbox. It was found that Linear Discriminant Analysis (LDA) functions best. Speed prediction was also done and values were determined through graphs. It was inferred that this categorizer efficiently translated all Sadhu words to Cholit precisely and in well-structured way. Therefore, Sadhu will not remain a complex language in this decade.

KEYWORDS: *Cholit, Inverse Data Frequency, Linear Discriminant Analysis, Principal Component Analysis, Sadhu, Term Frequency, Machine Learning*

How to cite this paper: Nakib Aman Turzo | Pritom Sarker | Biplob Kumar "Interpretation of Sadhu into Cholit Bhasha by Cataloguing and Translation System" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-3, April 2020, pp.1123-1130, URL: www.ijtsrd.com/papers/ijtsrd30792.pdf

IJTSRD30792

I. INTRODUCTION

Sadhu bhasha is a bygone ornate register related to Bengali vernacular, which is noteworthy used in the course of Bengali Renaissance from 19th to 20th century. It's different in its verb form, vocabulary and it's comprised of Sanskarit or tasama. It was exercised as penmanship unlike Cholito-vasha, which is unpretentious but used in longhand and also in verbalized form. The two types mentioned comes under diglossia. Most writings are carried out in Cholit bhasha. Areas of Bangladesh like Chittagong bears very superficial resemblance to cholit Bangla. In colonial era, Sadhu-vasha was exercised in formal dockets and licit papers though it's outworn in current era. Sadhu bhasha owes its origin to the literature by intellectuals of Gour. On the account of this this language is called Sadhu Gouriyo Bhasha. Cholit bhasha is more speaker friendly. For Bengali speakers Cholit bhasha is the most common bond of understanding and communication. In this paper, we worked on distinguishing a Bengali sentence whether it belongs to Sadhu-Vasha or Cholito-Vasha. This effort is first of its kind in metamorphosis of Sadhu vashha to Cholito. It may lead to creation of a software which can automatically detect if the sentence is in Sadhu or Cholito vasha and can translate the sentence to either language. The aim of this work is to familiarize the present generation of Bangladesh to classic literature by conversion of Sadhu to Cholitu and to translate the ancient legal dockets written in Sadhu to Cholito Vasha.

II. LITERATURE REVIEW

Classifiers reveal differences in grammar but not in cognition. Cantonese utilize over five sortal classifiers than Mandarian. Forty percent of nouns appear without classifier and 18% of Cantonese and 3% of Mandarian take a sortal [1].

In Mandarian and Cantonese, composition of an NP may be consists of just a classifier using semantic criteria to override their synctactic distributor [2].

Machine translation is a significant part of Natural Language Processing for conversion of one language to another. Translation consists of language model, translation model and a decoder. A statistical machine translation system was developed t translate English to Hindi. The model is developed by making use of software in Linux environment. [3].

Speech and language processing systems can be categorized according to predefined linguistic information use and is data driven and it made use of machine learning methods to automatically extract and process relevant units of information are indexed as appropriate. Therefore, an idea was exploited using ALISP (Automatic Language Independent Speech Processing) approach, with particularly focusing speech processing [4].

In a research it was shown that problem with many speech understanding systems was the context free grammar and augmented phrase structure grammars are very demanding

computationally. Finite state grammars are efficient but can't represent the relation of sentence meaning. It was described how language analysis can be tightly coupled by developing an APSG for analysis of component and deriving automatically. Using this technique efficient translation system was built that is fast compared to others [5].

In another research the integration of natural language and speech processing in Phi DM-Dialog and its cost-based scheme of ambiguity resolution were discussed. The simultaneous interpretation capability was made possible by an incremental parsing and generation algorithm [6].

Language conversion is toughest task and a case study was done for this trade-off. This included translation of client's system in proprietary language into programming languages. Various factors were considered that affect automation level of language conversion [7].

In 1996 CJK Dictionary Publishing Society launched an investigative project for the issues in depth and for making an elaborative simplified Chinese and traditional Chinese data base with 100% accuracy by collaborating with Basis Technology in developing sophisticated segmentation [8].

In few studies speech to text conversion of words were done for integrating people with hearing impairments. A software was developed to aid human being through correctness of pronunciation using English phonetics. This software helps in recognition of potential in English hearing [9].

An introduction of generic method for converting a written Egyptian colloquial sentence to diacritized Modern Standard Arabic (MSA) sentence which could easily be extended to be applied to other dialects of Arabic which could easily be applied to other dialects. A lexical acquisition of colloquial Arabic was done which is used to convert written Egyptian Arabic to MSA [10].

A system was also developed in this regard which recognizes two speakers in each of Spanish and English and was limited o 400 words. Speech recognition and language analysis are tightly coupled by using the same language model [11].

In a research by using neural network conversion of text written in Hindi to speech was done which has many applications in daily life for blind. It is also used for educating students. The document containing Hindi was used as input and neural network was used for character recognition [12].

Grammatical errors were quite restricted in variability and function in historical periods of English. In 19th and 20th century they become more productive accompanied by major extensions in function, variants and range of lexical association [13].

A Graphical User Interface has been designed for conversion of Hindi text to speech in java Swings because it consists of different languages spoken in different areas [14].

Recently progresses were made in speech synthesis has produced synthesizers with very high intelligibility but the naturalness and sound quality is still a problem. However, its quality has reached an adequate level for many applications [15].

There are many researches also aimed at recognition accuracy of speech with embedded spelled letter sequences. Different methods got proposed to localize spelled letter segments and reclassify them with a specialized letter recognizer [16].

Development report was prepared for translator software which partially offsets the absence of educational tools that hearing impaired, need for communication. For developing written language skills this tool could be used [17].

For converting words into triplets Software system converts between graphemes and phonemes using lexicon-based, rule based and data driven techniques. A shotgun integrate these techniques in a hybrid system and adds linguistic and educational information about phonemes and graphemes [18].

An online speech to text engine was developed for transfer of speech into written language in real time and it required special techniques [19].

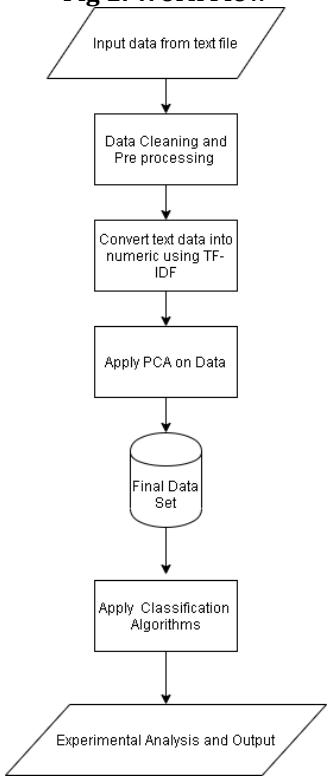
Examination of translation dilemmas was done in qualitative research. The medium of spoken and written language was critically challenged by taking into account the implications of similar problems. Centering translation and how it's dealt with issues raised by representation that would be concern for all researchers [20].

III. METHODOLOGY

Literature books were being used to gather all data regarding Sadhu and Cholit related sentences. Sum total of 2483 sadhu sentences from five significant literatures and 3508 cholit sentences from 6 important literature works were taken into account for this task.

The methodological steps used are as follows:
First we amassed a .txt file literature and then got well defined sentences from the literature. From each of the sentence we conjectured stop word. Then text sentence data is being metamorphosed to numeric data by utilizing TF-IDF. Final data set is obtained by application of PCA on data by using MATLAB and Python a variety of machine learning algorithms on the information set. At the ending point through analytical approach inspection is being done.

Fig 1: Work Flow



1. Data Clean

We have non- English (which got filtered out before or after processing of natural language data) in our set of information. All the non-English words got axed from it by us. Natural Language Toolkit (NLTK) information center of python is being used for this purpose. We have all of the sentences in non-English in our information set. Ergo, after the moping through the process, on the norm we got 1983 data set. As far as numeric categorization is concerned Sadhu is dubbed as numeric 0 and cholit is categorized as numeric 1.

2. Term Frequency–Inverse Document Frequency

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

3. Term Frequency (TF)

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

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

4. Inverse Data Frequency (IDF)

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

$$idf(w) = \log\left(\frac{N}{df_t}\right)$$

TF-IDF is simply the TF multiplied by IDF

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Our most work is being done from Scikit-Learn which is TF-DF Vectorizer's class. Our text data is taken by it and converted to numeric information set. After this conversion, our data has 3394 features. We have so many less important features we can do features extraction using PCA.

5. Principal Component Analysis

A new coordinate system is being metamorphosed from data through orthogonal linear transformation so that each coordinate has greatest variance by scalar projection of data in an ordered way and so on. This is called principal component analysis. Principal component analysis is a class of Scikit-learn. Higher variance comes to lie in first coordinate which is called first principal component and the lower variance in second coordinate. Our information set has 1678 traits after application of principal component analysis. When applications of dimensions of principal component analysis got reduced and the data quality got lost.

In case of principal quality analysis, 95% caliber of data was being maintained. 95% of the quality of real data was preserved by setting value of 'n' components as 0.95. Our latest data has 1678 characteristics after application of principal component analysis.

Data Set Prior To TF-IDF and PCA

'তিনি সাহিত্যিক সুতরং টাকার কথা তুলিয়া তাঁহার সারস্বত সাধনার অমর্যাদা করিতে চাই না	Sadhu
'[হায় রে পোড়াকপালে সাহিত্যিক৷] কিন্তু যদি তিনি নাম ধাম বদল করিয়া এই হীরা হরণের গল্পটা লিখিতে পরেন তাহা হইলে ত	Sadhu
'শ্রদ্ধা ও নমস্কার গ্রহণ করিবেন	Sadhu
'ভভারহেড 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 এবংগ্যালারি	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

Processed Data after TF-IDF and PCA

Fig 3

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

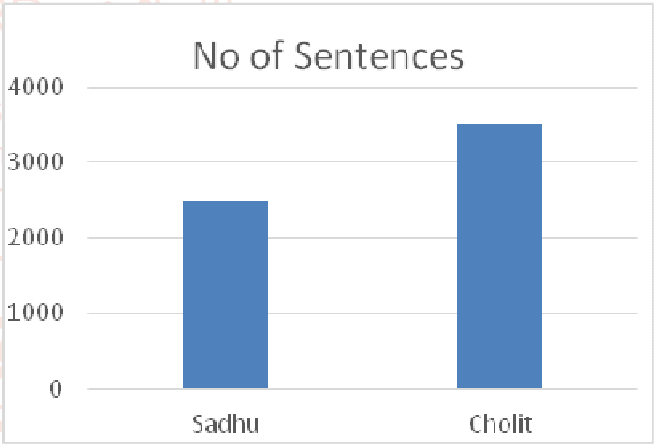


Fig 4

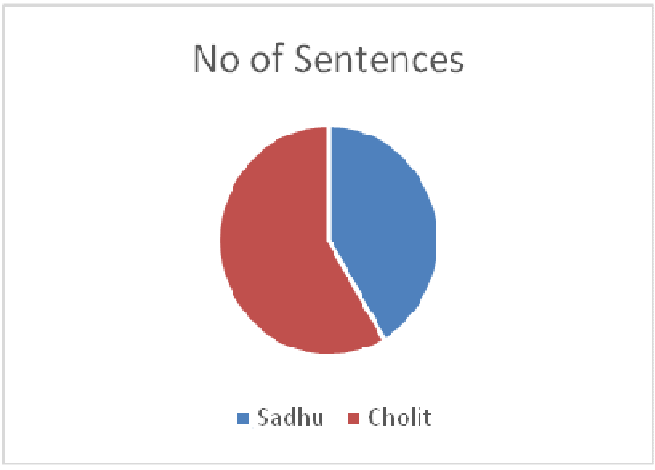


Fig 5

IV. RESULTS AND EXPERIMENTAL ANALYSIS

After implementing dataset in MATLAB results and factors for total misclassification of top 4 classifiers are as follows:

Classifier	Training Time	Total Misclassification Cost	Accuracy
Linear SVM	110.55	922	69.2%
Cosine KNN	72.37	912	69.6%
Boosted Trees	590.12	912	69.2%
Subspace Discriminant	580.69	939	68.7%

Table 1

Prediction speed graph showing Linear SVM has the fastest prediction speed and subspace discriminant being the slowest one. Naïve Bayes, tress classifiers were also used but discarded due to poor accuracy.

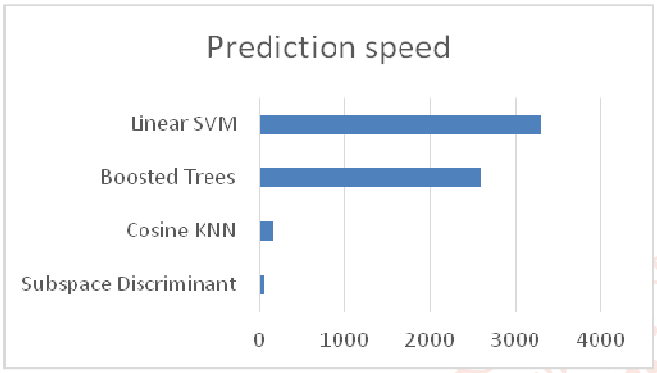


Fig 6

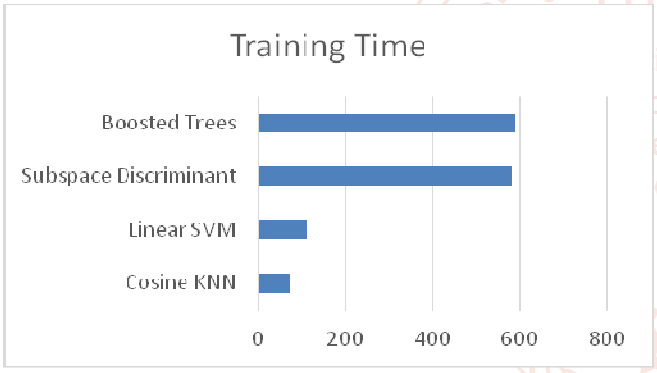


Fig 7

Cosine KNN has the fastest training time followed by linear svm. Ensemble classifiers like Boosted trees and subspace discriminant were much slower.

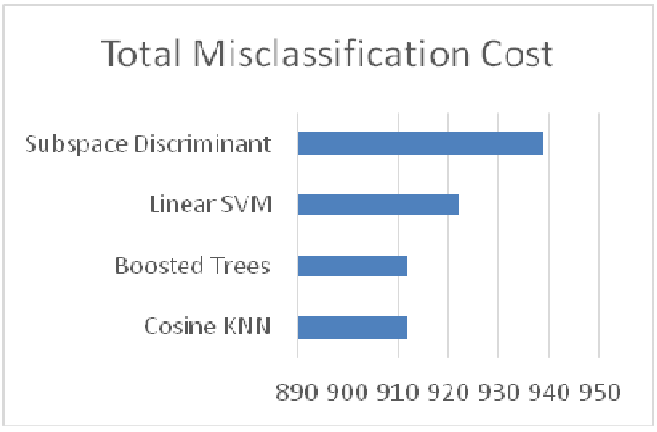


Fig 8

Cosine KNN and Boosted trees consumed to least amount of misclassification cost. Subspace discriminant had the most misclassification cost.

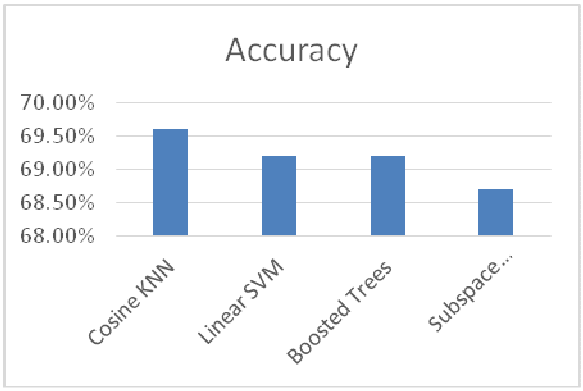


Fig 9

Cosine KNN gave the highest accuracy followed by Linear SVM.

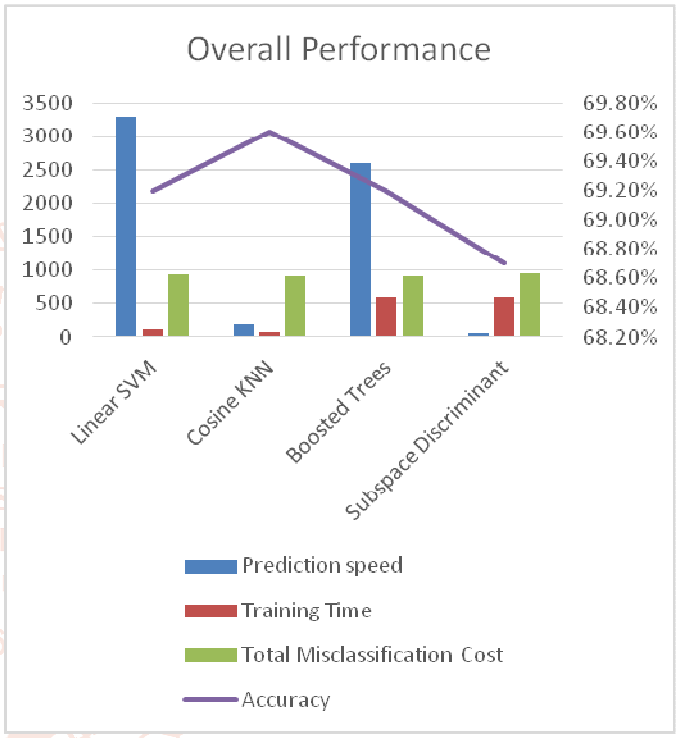
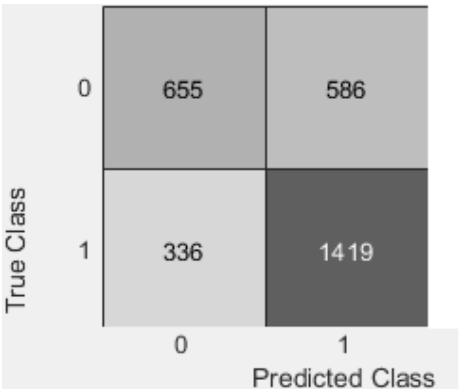


Fig 10

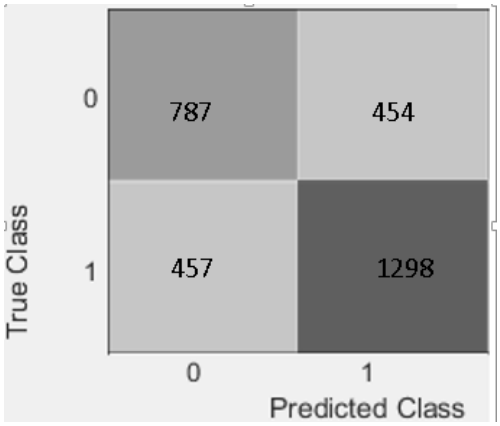
Though Cosine KNN performs the best and is ahead of others but it has the prediction speed lesser then others. The slot of 2nd accurate prediction speed is being taken by linear SVM which has low cost of altogether misclassification.

So in case of implementation of MATLAB and its optimization linear SVM is considered to be the best for classifying sadhu and cholit sentence.

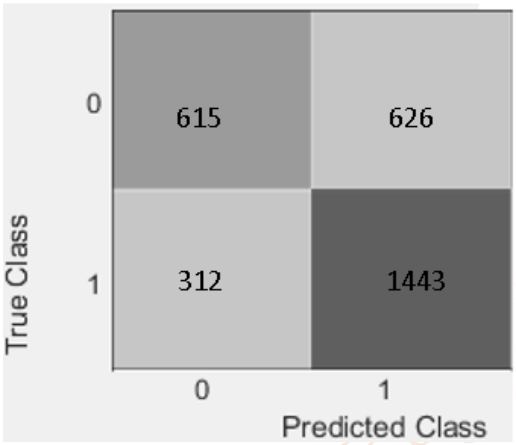
Confusion Matrix of Four Classifiers



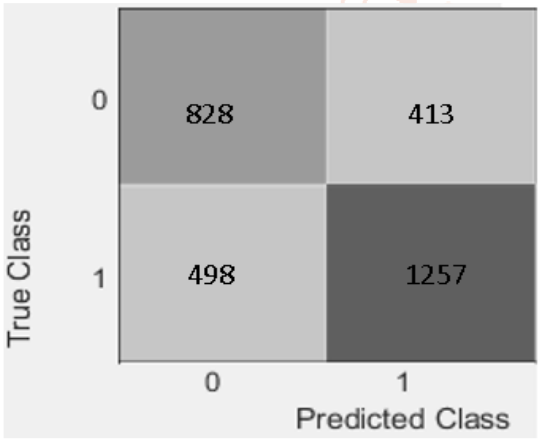
Linear SVM



Cosine KNN

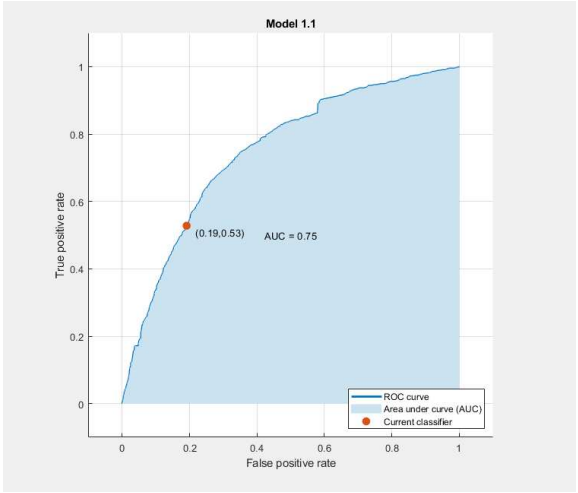


Subspace discriminant

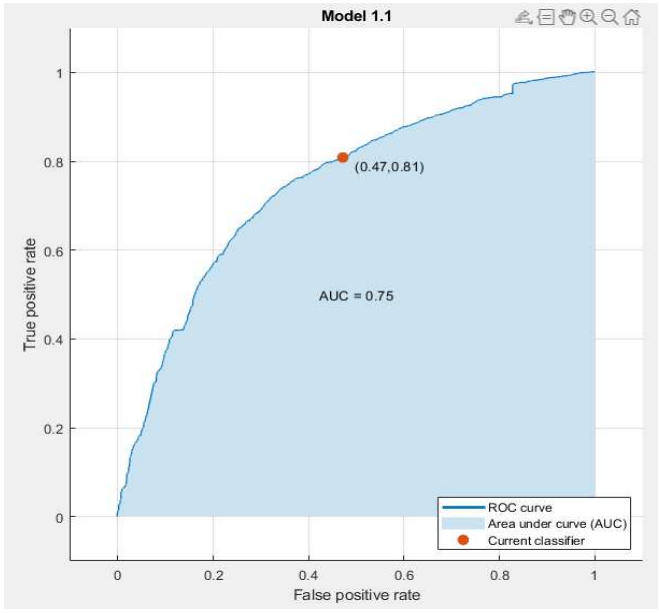


Boosted Trees

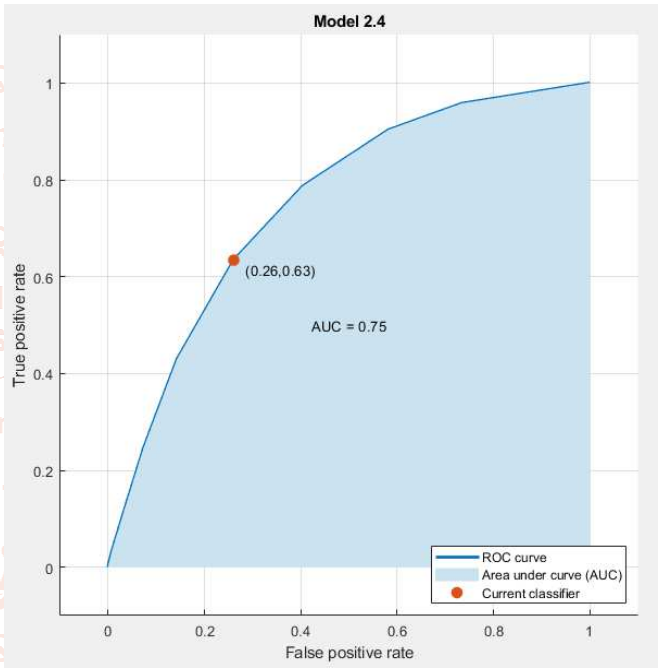
ROC Curve of Various Classifiers



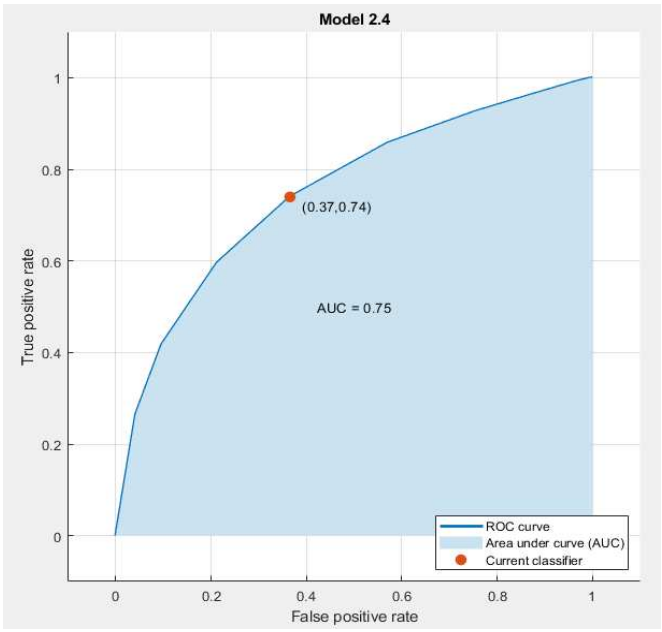
Linear SVM (Sadhu Sentence)



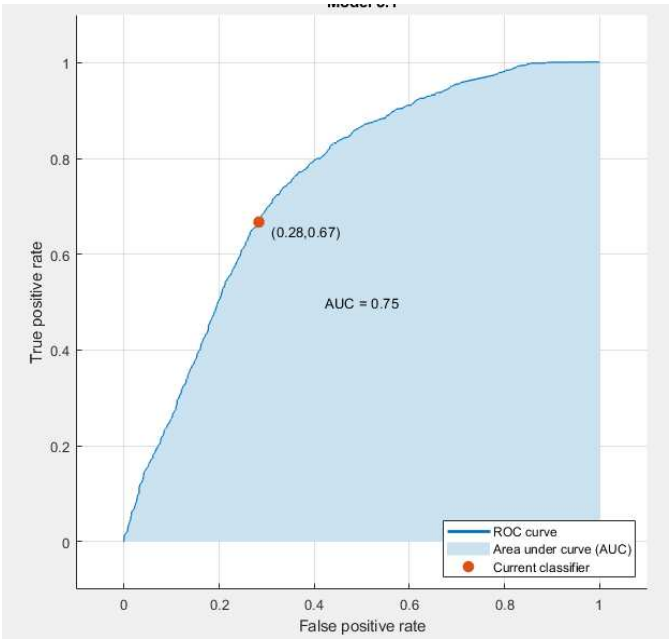
Linear SVM (Cholit Sentence)



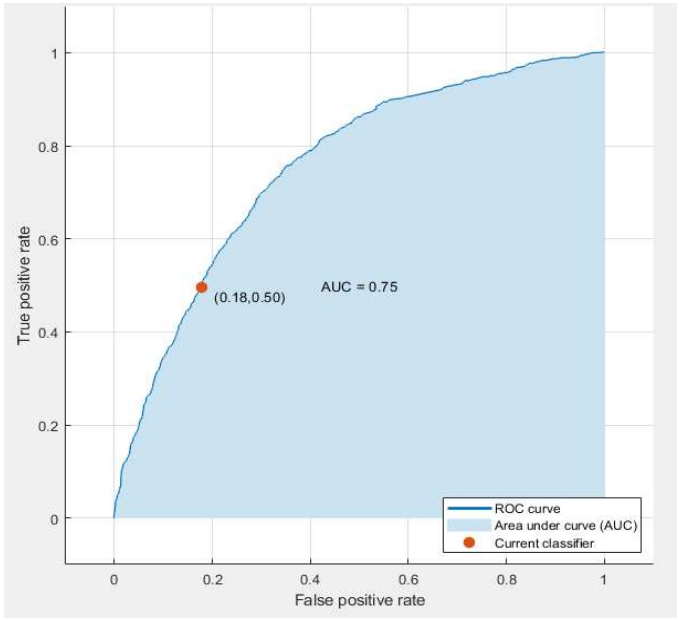
Cosine KNN (Sadhu Sentence)



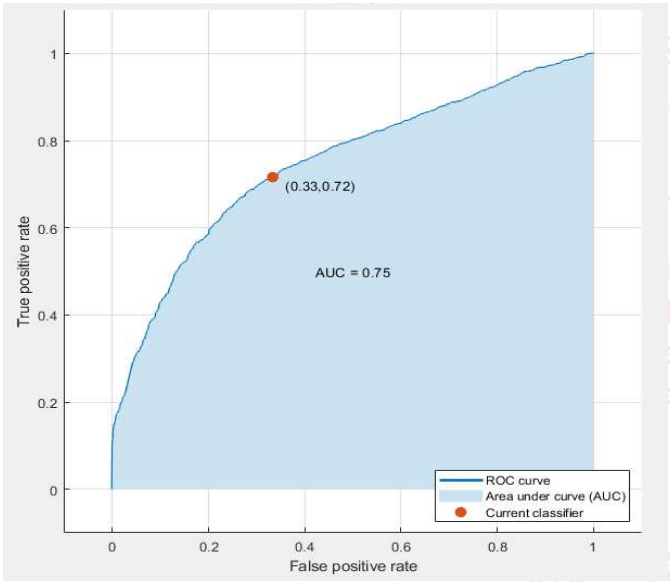
Cosine KNN (Cholit Sentence)



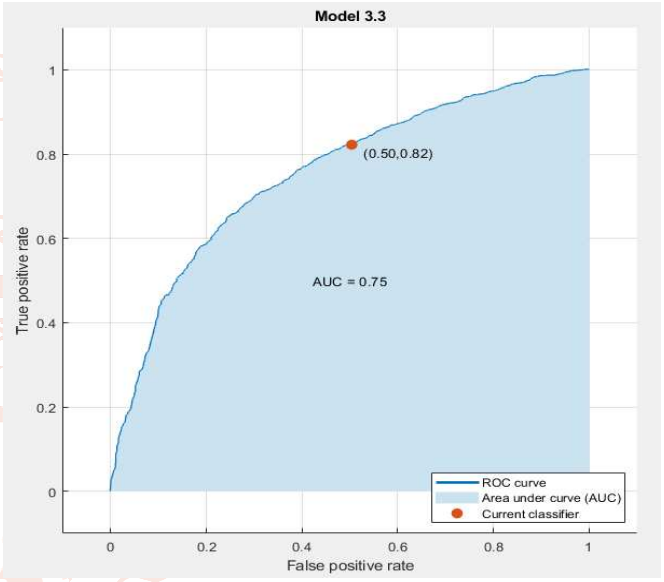
Boosted Trees (Sadhu Sentence)



Subspace discriminant (Sadhu Sentence)



Boosted Trees (Cholit Sentence)



Subspace discriminant (Cholit Sentence)

For ROC curves the steeper the curve the better the output. We get much steeper curve in linear SVM and cosine KNN.

Results after Dataset Implementation in Python

We used 15 algorithms of classification. For this operation we utilized scikit learn library. We have chosen five best models best on the cross validation score by doing it about 10 folds.

Accuracy Chart

	Accuracy	Recall	precision	F1	Kappa
Logistic Regression (python)	73.83%	0.7448	0.7966	0.7691	0.4675
SVM (Linear) (python)	74.04%	0.7444	0.7993	0.7703	0.4726
Ridge Classifier (python)	72.01%	0.696	0.8009	0.7441	0.4385
Linear discriminant analysis (python)	75.07%	0.7945	0.783	0.7885	0.4848
AdaBoost (python)	72.11%	0.7924	0.7469	0.7689	0.418

Table 3

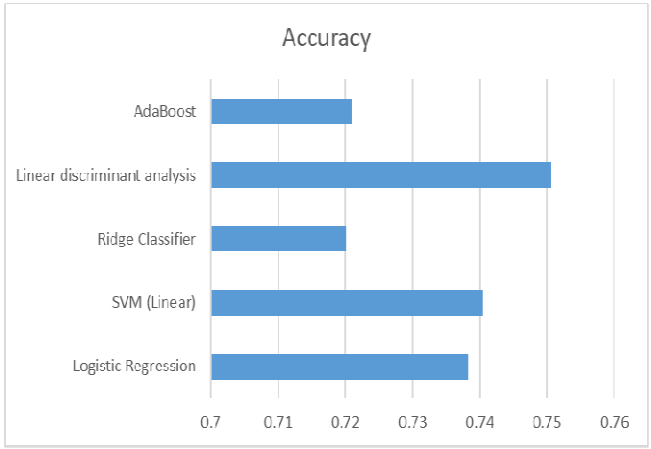


Fig 11

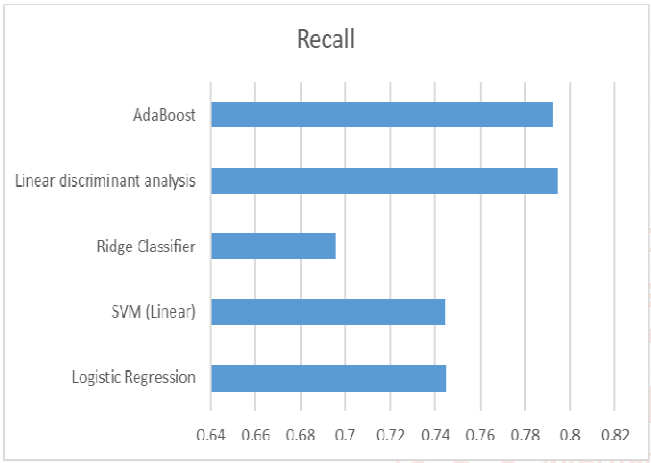


Fig 12

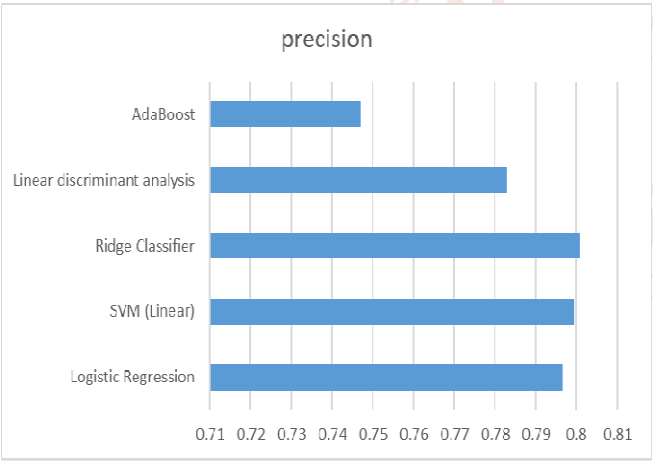


Fig 13

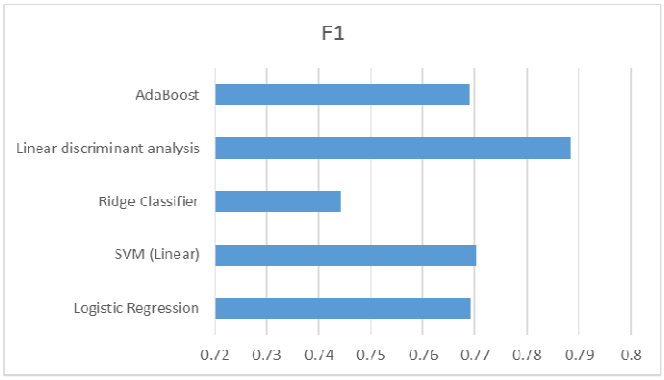


Fig 14

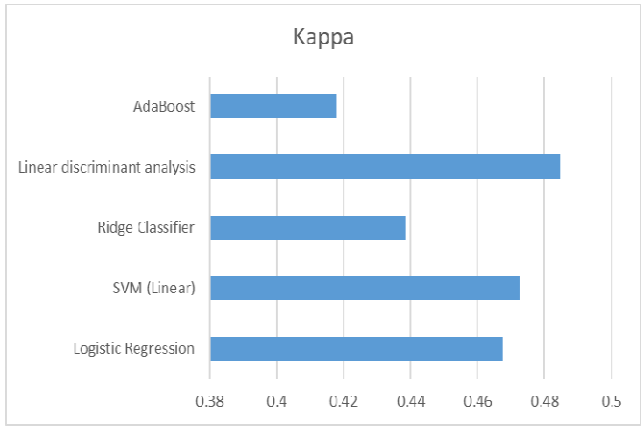


Fig 15

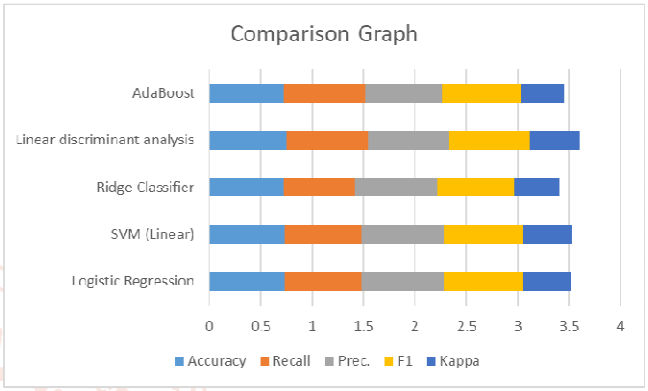
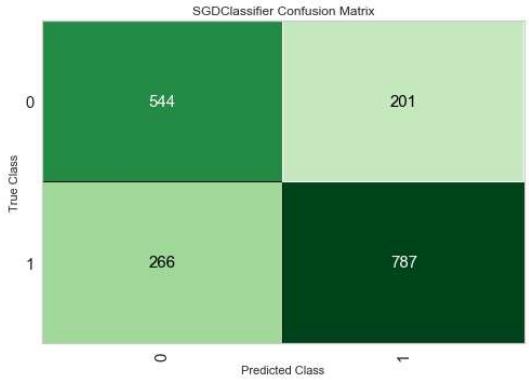


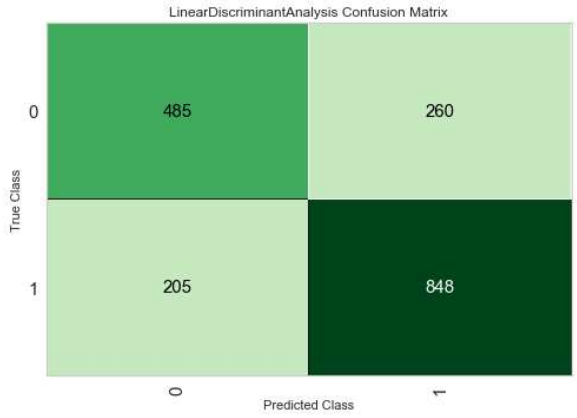
Fig 16

Among all classifiers linear discriminant analysis functions best as depicted in graph. In case of LDA, it is expressed as dependent variable as a linear combination of other features or measurements and has 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 is also resembled to them. When for each observation independent variables are continuous quantities DA also works there. Discriminant correspondence analysis is equivalent technique to categorical independent variables.

LDA has a close relation with SVM. For distinctively classifying the data point, the objective associated to support vector machine algorithm is to get a hyperplane in an N-dimensional space (N= Number of Features). There are two hyperplanes possible that could be selected to separate two distinctive classes of data points. The basic goal of our project is finding a plane that would have maximum margin, i.e. the maximum distance between data points of both classes. Future data points can be categorized by increasing the margin distance and provides reinforcement.

We can look at their confusion matrix for better understanding.





From the confusion matrix, we can clearly say that is high for the support vector machine but recall is high for the LDA.

V. CONCLUSIONS

As we consider whole algorithm, the precise results were given by LDA. This categorizer assists in classifying languages like Sadhu and Cholit. Sadhu being the most common language in past and also Bangladeshi literature is being enriched with it that is why most of the novels are written in Sadhu language. Sadhu is not in use in the present generation so for next step Sadhu is converted to Cholit so that people find ease in reading old era novels.

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."," in nternational 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] A. l. o. o. p. B. J. W. C. D.RileyAlejandroMacarrón, "A spoken language translator for restricted-domain context-free languages," Elsevier B.V., 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 text to speech conversion for Hindi language."," International Journal of Science and Research (IJSR), 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, Pittsburgh, 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, Gujarat, 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.