CATEGORIZATION AND TRANSLATION OPERATING SYSTEM'S ASSISTANCE IN EXPLICATION OF DIFFERENT BANGLADESHI ACCENTS

Article · June 2020

CITATIONS
0

READS
24

4 authors:

Nakib Aman Turzo
Varendra University
16 PUBLICATIONS **1** CITATION
SEE PROFILE

Pritom Sarker
Varendra University
10 PUBLICATIONS **1** CITATION
SEE PROFILE

Biplob Kumar
Varendra University
6 PUBLICATIONS **1** CITATION
SEE PROFILE

Niloy Kumar Shaha
University of Rajshahi
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

European Journal of Computer Science and Information Technology
Vol.8, No.3, pp.31-45, June 2020
Published by *ECRTD- UK*
Print ISSN: 2054-0957 (Print), Online ISSN: 2054-0965 (Online)

CATEGORIZATION AND TRANSLATION OPERATING SYSTEM'S ASSISTANCE IN EXPLICATION OF DIFFERENT BANGLADESHI ACCENTS

Nakib Aman Turzo
Lecturer,
Department of Computer Science & Engineering,
Varendra University,
Rajshahi, Bangladesh
Email: nakibaman@gmail.com

Pritom Sarker
B.Sc. in CSE,
Department of Computer Science & Engineering,
Varendra University,
Rajshahi, Bangladesh
Email: me.pritom@gmail.com

Biplob Kumar
B.Sc. in CSE,
Department of Computer Science & Engineering,
Varendra University,
Rajshahi, Bangladesh
Email: kumarbiplob336@gmail.com

Niloy Kumar Shaha
B.Sc. in CSE,
Department of Computer Science & Engineering,
Varendra University,
Rajshahi, Bangladesh
Email: niloyshaha20@gmail.com

ABSTRACT: *National language of Bangladesh is Bengali and it's also the official language used frequently. Our paper's focal point was to categorize and differentiate West Bangla language or Bangladeshi Bangla accent in a Bengali sentence. We first amassed text from literature files. Then converted text sentence data to numeric data by using TF-IDF. After PCA application by MATLAB, final data set was being obtained. Our strategy for future will assist in developing an automatic software that detects if a sentence has been written in West Bangla or Bangladeshi Bangla and then it will do translation from one to another form. Differences between both Bangladeshi accents is already so minimum that only native speaker can identify them distinctively. There was no data available previously for this study. This work denoted that as if languages seems to be same but are unique and different in their own way and depicts the identity of two geographically separated regions. The major output of this work paid heed on identification of the form of language frequently used today. Many other studies could be conducted, based on the results of our study, on the effects of Sanskrit and Foreign literature*

KEYWORDS: Bangladeshi Bangla, Inverse Data Frequency, Linear SVM, Principal Component Analysis, Python, Term Frequency, West Bangla

INTRODUCTION

Bangladesh's official and national language is Bengali with respect to Constitution's third article. 98% of Bangladeshis are fluent in Bengali as their first language. Bengali dialects are being classified in two dimensions i.e spoken vs. literary variations and prestige vs. regional variations. Spoken Bengali exhibits more variations than written one. Formal language including in speeches, news, announcements is in Cholit Bhasha. During Bengali standardization in late 19th and early 20th century, cultural elite mostly belong to regions like Kolkata, Hooghly, Howrah and Nadia. In both Bangladesh and west Bengal the standard today is based on West Central Dialect while the language has been standardized through to centuries of media and education with mostly speakers fluent in both their socio-geographical variety as well as the standard dialect used in the media. Differences in dialects are in three forms literary language vs. colloquial language, regional dialectvs. Standardized dialect and lexical variations. Dialectal names are originated from the districts where they belong. Standard form doesn't show much varieties across Bengali speaking areas of South Asia. Variations which are regional in spoken Bengali constitutes a dialect continuum.

Mostly speech differences occurs at a distance of few miles and have distinct forms among religious communitive vocabularies. Bengali Hindus tend to speak in Sanskritised Bengali while Bengali Muslims use Perso-Arabic. Western border dialects are spoken in the area known as Manbhumi. There are many more minor dialects as well including those spoken in the bordering districts of Purnea and Singhbum and among the tribals of eastern Bangladesh like Chakma and Hajong. Bengali's rich literature prior to 19th century was in rhymed verse. Writing system of Modern Bengali developed from an ancient Indian syllabary called Brahmi. Like all Brahmi scripts Bengali is being written from left to right with characters hanging from horizontal line. No distinction is present in upper and lower case letters.

LITERATURE REVIEW

Classifiers uncover contrasts in language yet not in cognizance. Cantonese use more than five sortal classifiers than Mandarian. 40% of things show up without classifier and 18% of Cantonese and 3% of Mandarian take a sortal [1].

Creation of a NP in Mandarian and Cantonese might be comprises of only a categorizer by using semantic measures to supersede their synctactic merchant [2].

Machine interpretation is a critical piece of Natural Language Processing (NLP) for transformation of one language to another. Interpretation comprises of language model, interpretation model and a decoder. A measurable machine interpretation framework was created to make an interpretation of English to Hindi. The model is created by utilizing programming in Linux condition [3].

Discourse and language preparing frameworks can be sorted by predefined etymological data use and is information driven and it utilized AI techniques to consequently concentrate and procedure applicable units of data are ordered as proper. In this way, a thought was misused utilizing ALISP (Automatic Language Independent Speech Processing) approach, with especially centering discourse handling [4].Issue with numerous discourse understanding frameworks was the setting free language structure and enlarged expression structure syntaxes are requesting computationally.

Limited state language structures are effective however can't speak to the connection of sentence meaning. It was portrayed how language investigation can be firmly coupled by building up an APSG for examination of part and determining naturally. Utilizing this strategy proficient interpretation framework was manufactured that is quick contrasted with others [5]. In another exploration the mix of regular language and discourse preparing in Phi DM-Dialog and its cost-based plan of equivocalness goals were talked about. The synchronous understanding ability was made conceivable by a steady parsing and age calculation [6].

Change of language is the hardest assignment and a contextual investigation was accomplished for this exchange off. This remembered interpretation of customer's framework for restrictive language into programming dialects. Various components were considered that influence robotization level of language transformation [7].

In 1996 CJK Dictionary Publishing Society began an analytical task for the issues top to bottom and for making an elaborative streamlined Chinese and conventional Chinese information base with 100% exactness by working together with Basis Technology in creating advanced division [8].

In hardly any investigations discourse to content of words transformation were accomplished for reconciliation of individuals with hearing weaknesses. Improvement of programming was to help individual through rightness of elocution utilizing English phonetics. This product helps in acknowledgment of potential in English hearing [9].

A presentation of nonexclusive technique for changing over a composed Egyptian everyday sentence to diacritized Modern Standard Arabic (MSA) sentence which could without much of a stretch be reached out to be applied to different vernaculars of Arabic which could undoubtedly be applied to different lingos. A lexical obtaining of informal Arabic was done which is utilized to change over composed Egyptian Arabic to MSA [10]. A framework was likewise evolved in such manner which perceives two speakers in every one of Spanish and English and was constrained o 400 words. Discourse acknowledgment and language examination are firmly coupled by utilizing a similar language model [11].

In an examination by utilizing neural system transformation of content written in Hindi to discourse was done which has numerous applications in everyday life for daze. It is likewise utilized for teaching understudies. The report containing Hindi was utilized as information and neural system was utilized for character acknowledgment [12].

There is limitation of syntactic blunders in inconstancy and capacity in verifiable times of English. In nineteenth and twentieth century they become increasingly beneficial joined by significant expansions in capacity, variations and scope of lexical affiliation [13].

For transformation of Hindi content to discourse in Java Swings a Graphical User Interface has been planned on the grounds that it comprises of various dialects spoken in various zones [14].

As of late advances were made in discourse union has delivered synthesizers with exceptionally high comprehensibility yet the expectation and sound quality is as yet an issue. Be that as it may, its quality has arrived at a sufficient level for some applications [15].

There are numerous looks into likewise focused on acknowledgment exactness of discourse with installed spelled letter groupings. Various techniques got proposed to limit spelled letter portions and rename them with a particular letter recognizer [16].

Improvement report was set up for interpreter programming which mostly counterbalances the nonattendance of instructive devices that conference weakened, requirement for correspondence. For creating composed language abilities this apparatus could be utilized [17].

For changing over words into triplets Software framework changes over among graphemes and phonemes utilizing vocabulary based, rule based and information driven strategies. A shotgun incorporate these strategies in a half and half framework and includes etymological and instructive data about phonemes and graphemes [18].

An online discourse to content motor was produced for move of discourse into composed language continuously and it required exceptional procedures [19].

Examination of interpretation situations was done in subjective research. Vehicle of composed and communicated in language was fundamentally tested by considering the ramifications of comparable issues. Interpretation as essential issue and how its managed issues raised by portrayal that would be worry for all analysts [20].

A similar work was also performed on differentiation and translation of Sadhu and Cholit language which was the basis of inter-conversion of other languages. Here Linear Discriminant Analysis performed best and speed prediction was also done. That is why Sadhu didn't remain complex language [21].

METHODOLOGY

Sum total of 28550 sentences were taken into account for this task. Altogether 10800 Bengali sentences from 8 distinctive literatures were being selected and 17750 Bengali sentences from West Bengal were picked from 10 distinct literatures.

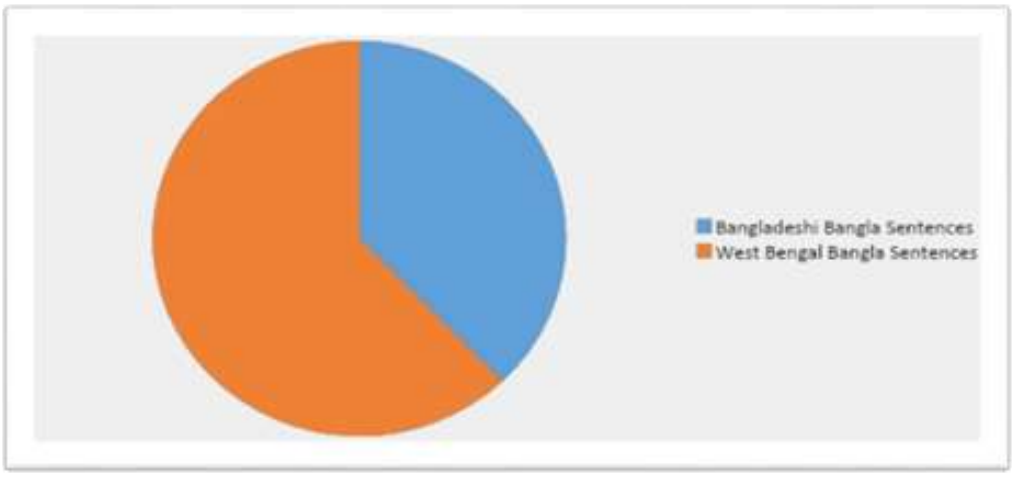

Fig 1

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.

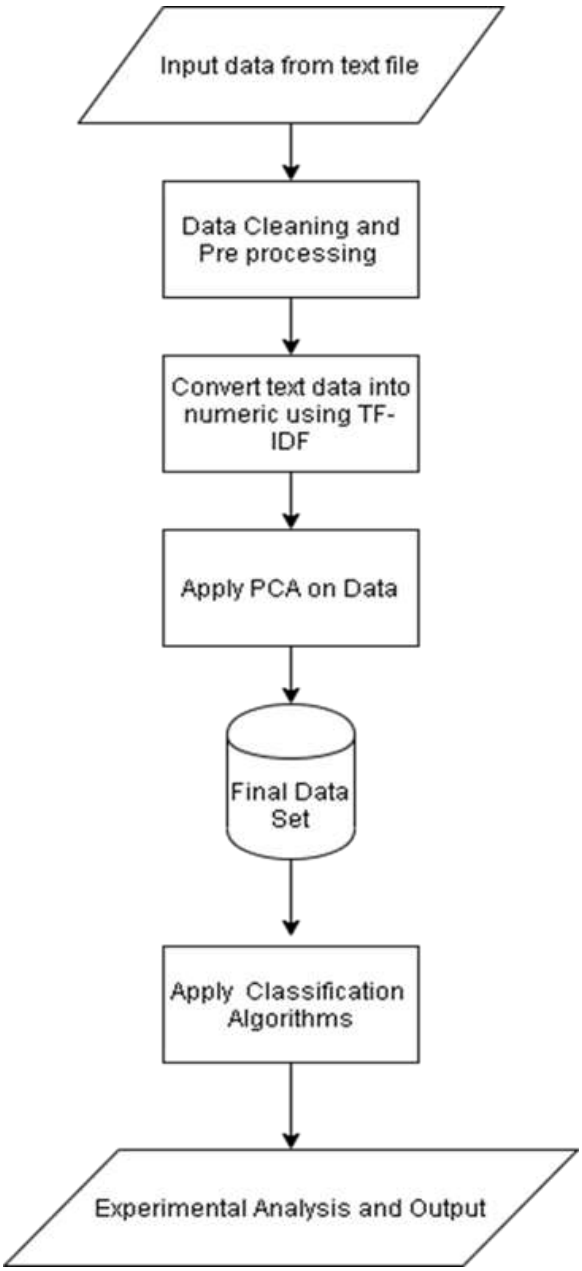

Fig 2: Work Flow

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.

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.

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.

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

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.

Principal Component Analysis

For pages other than the first page, start at the top of the page, and continue in double-column format. The two columns on the last page should be as close to equal length as possible. 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.

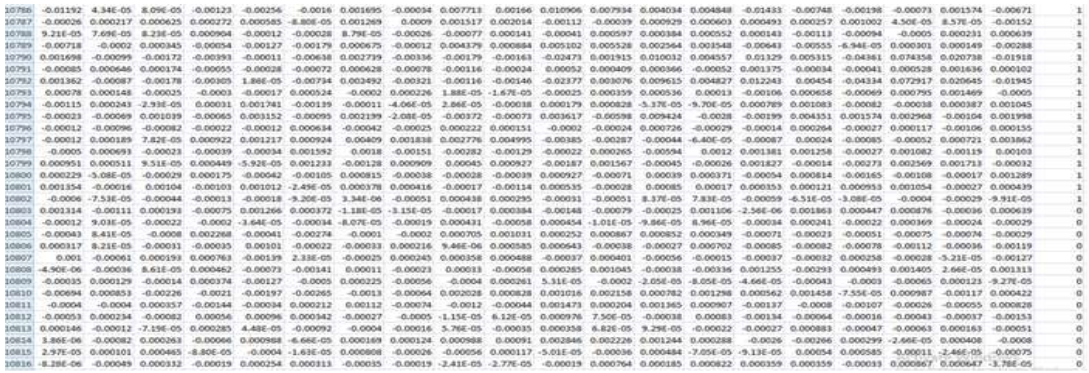

Fig 3

There are 1194 different sides of numeric data in which the last field denotes 1 for Bangla language of Bangladesh while 0 for Bangla language of West Bengal.

RESULTS AND EXPERIMENTAL ANALYSIS

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

Table 1

Classifier Name	Prediction speed	Training Time	Total Misclassification Cost	Accuracy
Linear SVM	1200	2715.7	3385	76.3%
Quadratic SVM	46	3564.7	3324	76.7%
Medium Gaussian SVM	58	1775.7	3448	75.8%
Bagged Trees	1600	740.01	4041	71.7%
Subspace Discriminant	110	1057.9	3364	76.4%

Bagged Trees depiction through prediction speed graph has maximum speed of anticipation while the Quadratic SVM has the lowest anticipation speed. Categorizers like Naïve Bayes and Decision Tree are utilized but ameliorated due to less precision.

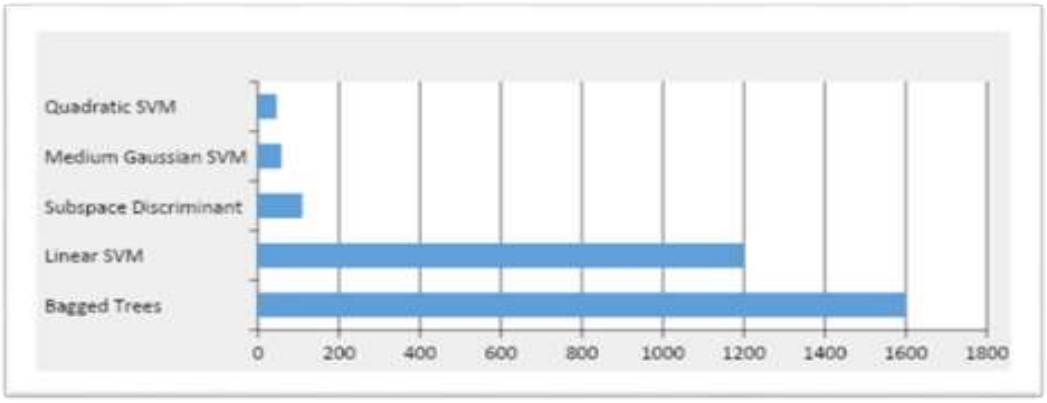

Fig 4

Subspace Discriminant comes after Bagged Tree which has highest training time. SVM categorizers are much slow in this regard.

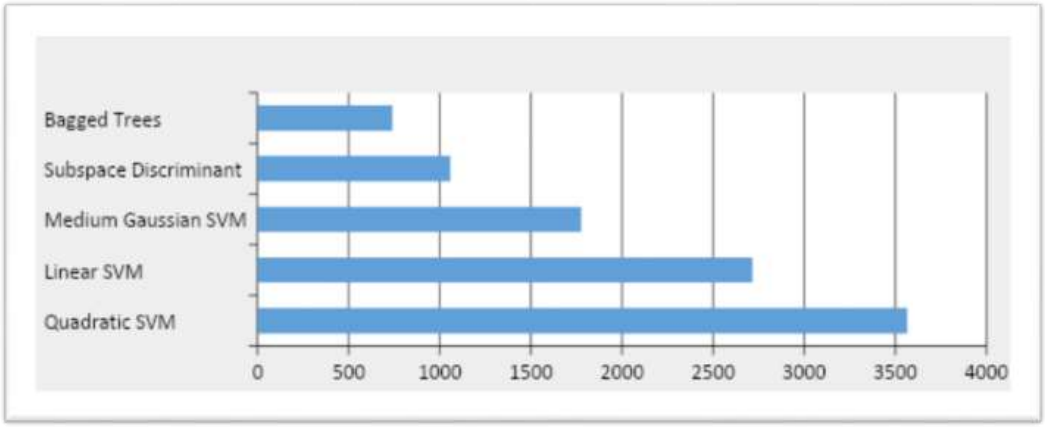

Fig 5

Fascinating features depicted here is total misclassification cost. Bagged tree which has previously highest value is the only categorizer lowest here while the others are similar.

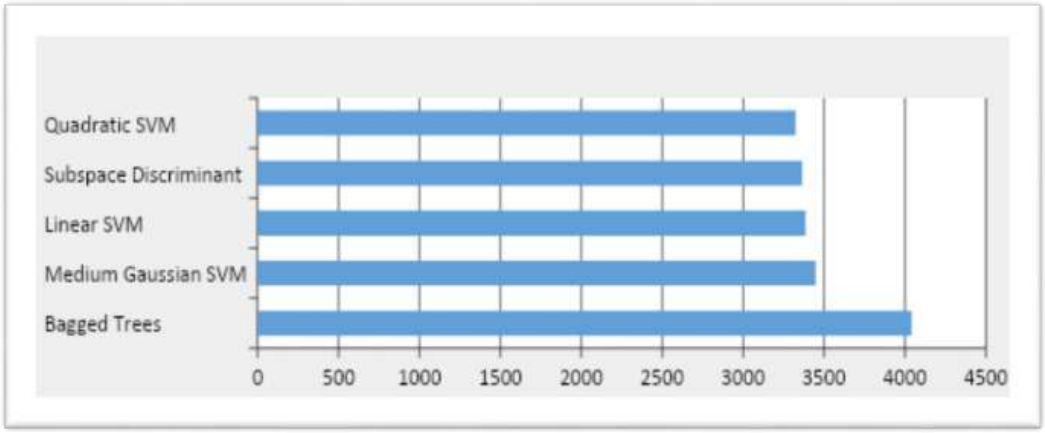

Fig 6

Quadratic SVM gave the highest accuracy followed by Subspace Discriminant.

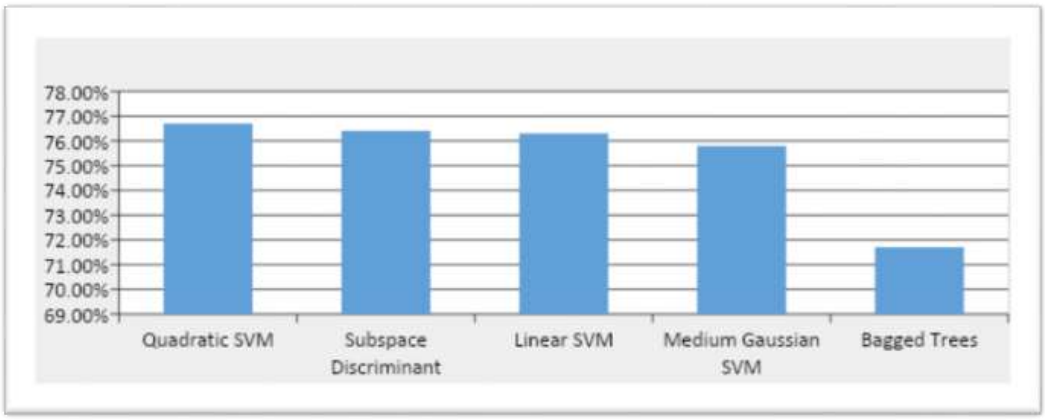

Fig 7

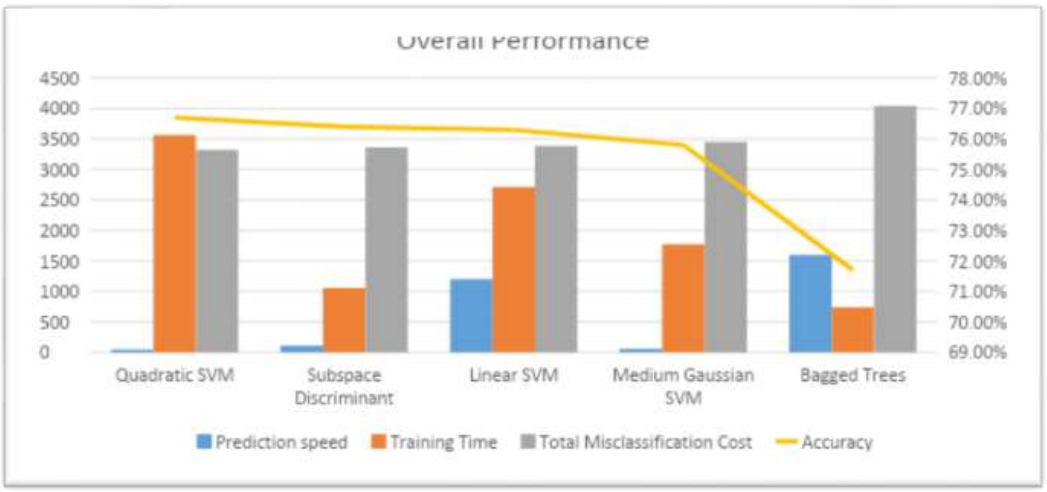

Fig 8

It can be observed from the functions that Quadratic SVM is highly precise and has high training time. Linnear SVM is at 3rd rank with high precision. So after optimization in MATLAB, Linear SVM is beneficial for categorizing Bangladeshi Bangla and West Bengali Bangla.

Here are the preset values used for the classifiers –

Table 2

Quadratic SVM	**Linear SVM**	**Medium Gaussian SVM**	**Boosted Trees**	**Subspace Discriminant**
Model Type	**Model Type**	**Model Type**	**Model Type**	**Model Type**
Preset: Quadratic SVM	Preset: Linear SVM	Preset: Gaussian SVM	Preset: Bagged Trees	Preset: Subspace Discriminate Kernel
Kernel function: Quadratic	Kernel function: Linear	Kernel function: Gaussian	Kernel function: Bag	Kernel function: Subspace
Kernel scale: Automatic	Kernel scale: Automatic	Kernel scale: 35	Learner Type: Decision tree	Learner Type: Discriminant
Box constraint level: 1	Box constraint level: 1	Box constraint level: 1	Maximum number of split: 28551	Maximum number of split: 30
Multiclass method: One-vs-One	Multiclass method: One-vs-One	Multiclass method: One-vs-One	Number of learner: 30	Number of learner: 597
Standardize data: true	Standardize data: true	Standardize data: true		

Confusion Matrix of 5 classifiers:

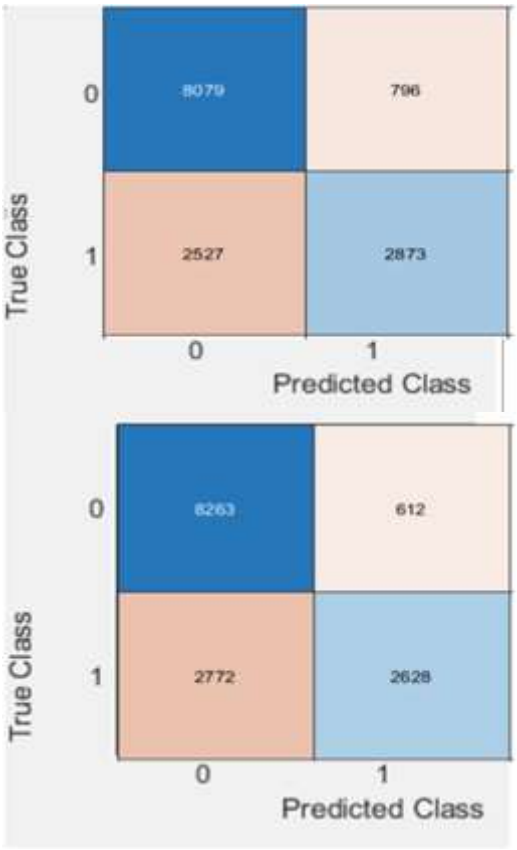

Fig 9: Quadratic SVM

Fig 10: Linear SVM

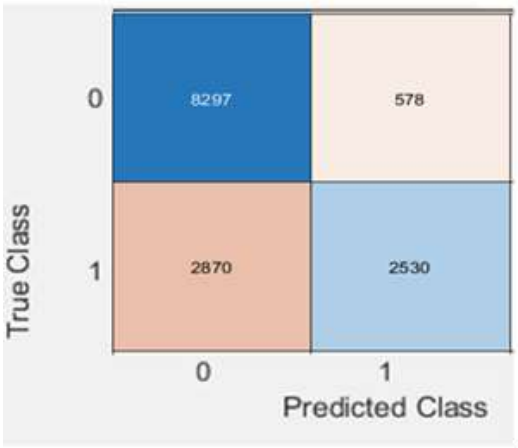

Fig 11: Medium Gaussian SVM

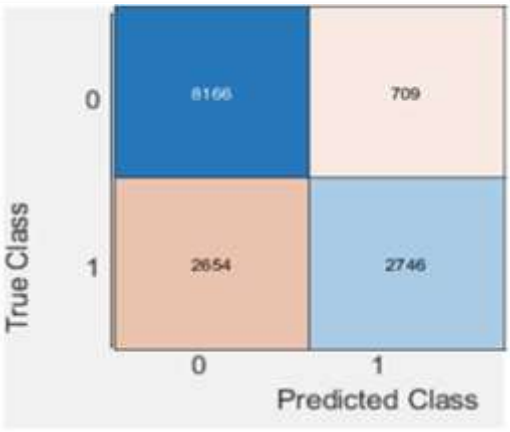

Fig 12: Boosted Trees

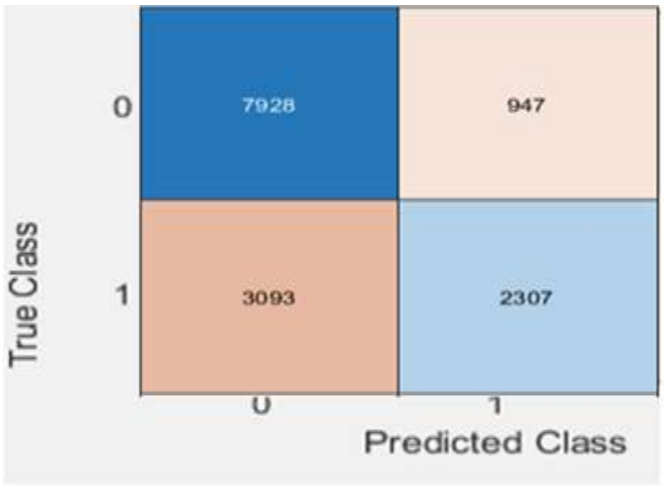

Fig 13: Subspace discriminant

ROC curve of various classifiers:

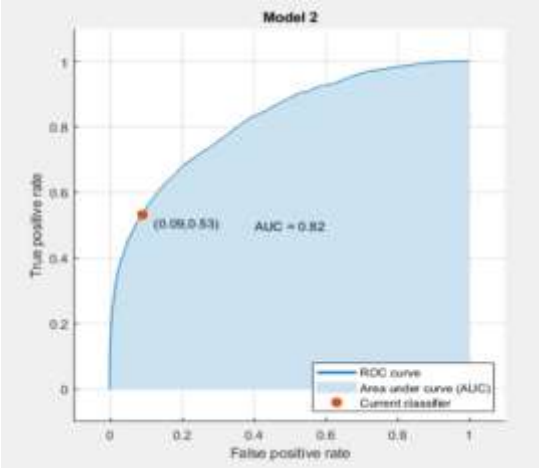

Fig 14: Quadratic SVM (Bangladeshi Bangla)

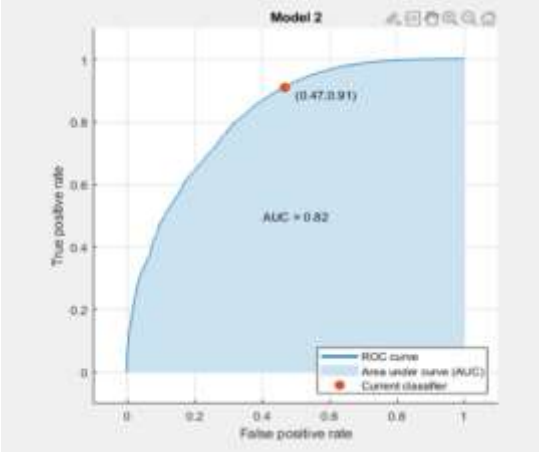

Fig 15: Quadratic SVM (West Bengal Bangla)

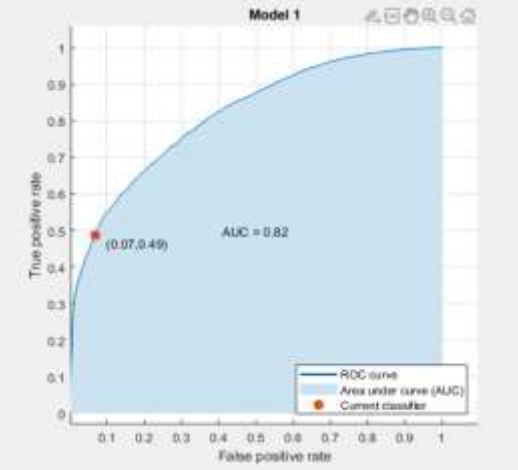

Fig 16: Linear SVM (Bangladeshi Bangla)

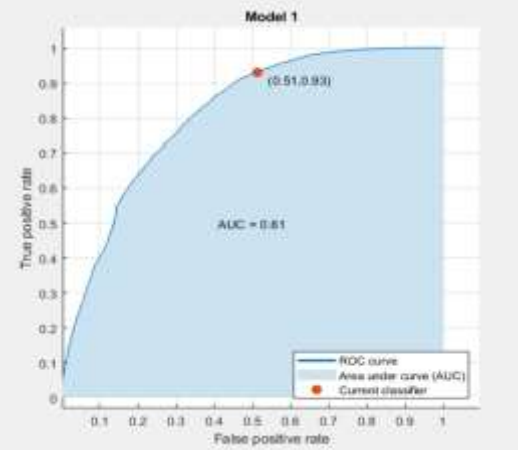

Fig 17: Linear SVM (West Bengal Bangla)

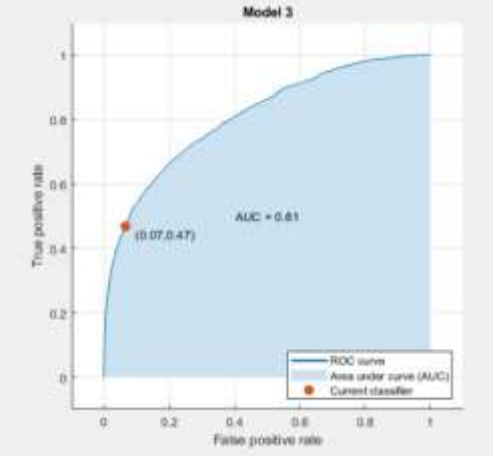

Fig 18: Medium Gaussian SVM (Bangladeshi Bangla)

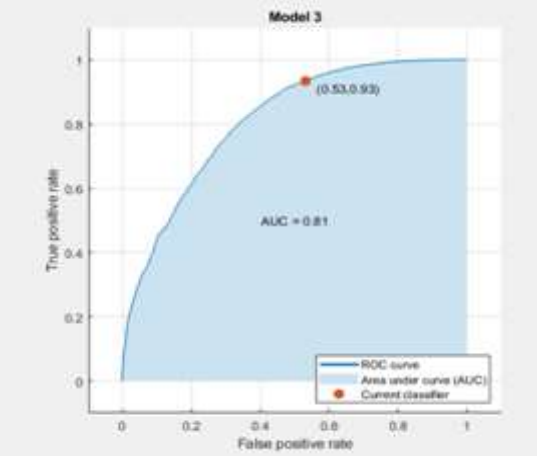

Fig 19: Medium Gaussian SVM (West Bengal Bangla)

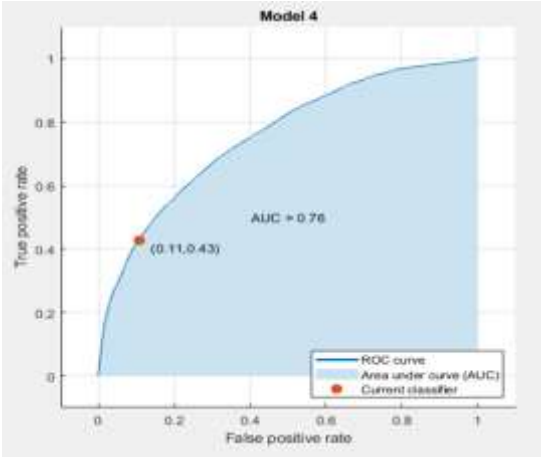

Fig 20: Boosted Trees (Bangladeshi Bangla)

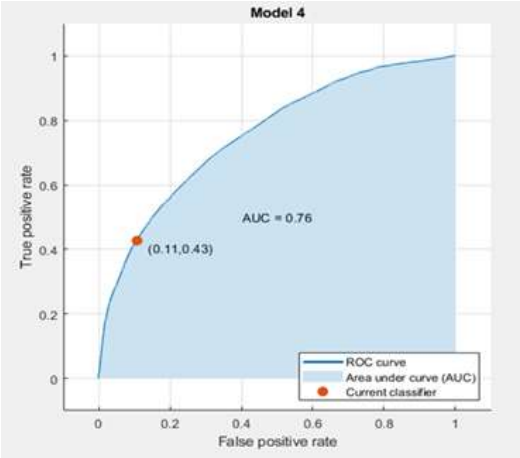

Fig 21: Boosted Trees (West Bengal Bangla)

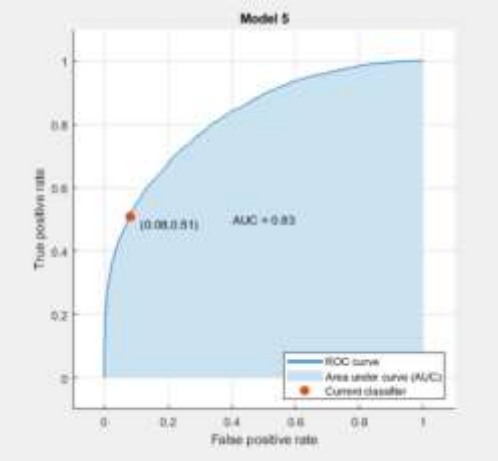

Fig 22: Subspace discriminant (Bangladeshi Bangla)

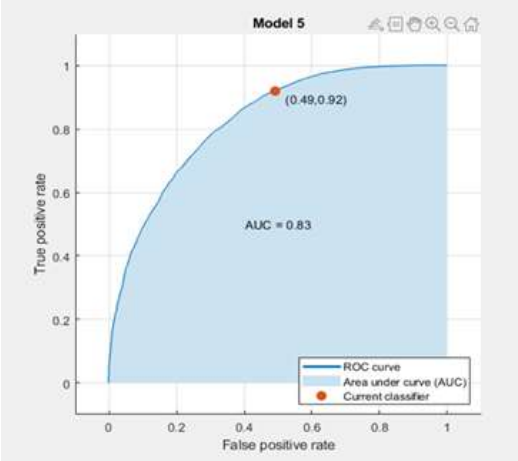

Fig 23: Subspace discriminant (West Bengal Bangla)

In case of ROC curves output is directly related to steepness and in linear SVM we get much steeper curve value. Thus Linear SVM depicts the best functionality among categorizing Bangladeshi Bangla and West Bengal Bangla classification.

CONCLUSION

As we consider whole algorithm, the precise results were given by Linear SVM and gives the expected outcomes. This categorizer assists in classifying languages like Bangla of Bangladesh and Bangla of Western Bengal. This categorizer would prove useful in classifying other accents too and the differentiation of languages will become much easier and understandable.

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.

[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, pp. 1123-1130, 2020.