Influence of Target Reader Background and Text Features on Text Readability in Bangla: A Computational Approach

Manjira Sinha
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
West Bengal, India
manjira@cse.iitkgp.ernet.in

Tirthankar Dasgupta
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
West Bengal, India
tirtha@cse.iitkgp.ernet.in

Anupam Basu
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
West Bengal, India
anupam@cse.iitkgp.ernet.in

Abstract

In this paper, we have studied the effect of two important factors influencing text readability in Bangla: the target reader and text properties. Accordingly, at first we have built a novel Bangla readability dataset of 135 documents annotated by 50 readers from two different backgrounds. We have identified 20 different features that can affect the readability of Bangla texts; the features were divided in two groups, namely, 'classic' and 'non-classic'. Preliminary correlation analysis reveals that text features have varying influence on the text hardness stated by the two groups. We have employed support vector machine (SVM) and support vector regression (SVR) techniques to model the reading difficulties of Bangla texts. In addition to developing different models targeted towards different type of readers, separate combinations of features were tested to evaluate their comparative contributions. Our study establishes that the perception of text difficulty varies largely with the background of the reader. To the best of our knowledge, no such work on text readability has been recorded earlier in Bangla.

1 Introduction

Readability of a text generally refers to how well a reader is able to comprehend the content of a text, through reading (Dale and Chall, 1948). Readability is a complex cognitive phenomenon where, the cognitive load of a text for a reader depends on both the characteristics of a text like, lexical choice, syntactic complexity, semantic complexity, discourse level complexity and on the background of the user. Several experiments have already established that readability of texts are quite language dependent and existing readability measures in English cannot directly be used to compute readability of other languages like, Bangla and Hindi (Sinha et al., 2012). Yet, compared to the numerous readability measures in English and other European languages(Benjamin, 2012), few initiatives have been taken to compute text readability in a Eastern Indo-Aryan language like Bangla or any other Indian languages which are structurally very different from many of their Indo-European cousins such as English, which is of West-Germanic descent (Sinha et al., 2012). One important factor that affects the readability of a text is the background of the respective reader. According to Dale (Dale, 1949), "The interpretation of the expressed thought is related more to the reader's informational background and motivations than to the internal evidences of the expressional facility of the author". Reader's background is a complex derivative of one's educational and socio-economic state. As per one of the pioneering works in readability by Dale and Chall (1949), the outcome of reading depends on many characteristics of the prospective readers including "reading abilities, interests, age, sex, intellectual

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 345–354, Dublin, Ireland, August 23-29 2014.

maturity, background of information etc." However, we do not know of any such investigations for Bangla text readability that have investigate the way background of a reader affect the readability of text. Such language specific study is needed as Bangla as a language is very different from English and the inapplicability of English readability formulae for Bangla text has already been established.

Considering the above issues as our motivation, in this paper we have developed models to predict reading difficulty of a Bangla document perceived according to different target reader groups. To categorize among different reader groups, we have considered age, education and socio-economic data as indicators of comprehension ability. In addition, we have also explored the impact of different types of text features on text comprehensibility in Bangla. However, development and evaluation of such model requires availability of well-annotated resources. To the best of our knowledge, no automatically accessible data annotated according to the reading difficulty level is available for Bangla. Therefore, we have developed a digital resource pool of Bangla text documents in Unicode encoding that can be used for various NLP tasks such as feature extraction, document analysis etc. Such a dataset is essential to analyze readability of text documents based on the target reader. Next, we have visualized the text readability problem from a machine learning perspective as a classification problem using support vector machines (SVM) and an estimation problem using support vector regression (SVR). Our study is based on a wide range of textual features, from the syntactic and lexical features of a text like, its average sentence length, average word length in terms of visual units, to discourse level features like, number of jukta-akshars (consonant conjuncts) , number of different parts of speeches, named entity and lexical relations (refer to section 3). Although regression analysis has been previously used to model the text readability in Bangla, reader group specific analysis and machine learning techniques like support vectors have not been used so far. We have considered two target reader groups namely Group-1(or Adult group) with average age of 23 Yrs and Group-2 (or minor's group) with average age of 15 Yrs.

The organization of the paper is as follows: section 2 presents a brief literature survey on existing readability metrics for English and Bangla; section 3 defines the features of a text considered in this study, and empirical data collection, section 4 discusses the experiment observations, the prediction techniques and presents the results and validations for the two techniques. Finally, section 5 offers conclusion and perspective.

2 Related Works

The quantitative analysis of text readability started with L.A. Sherman in 1880 (Sherman, 1893). Till date, English and other languages have got over 200 readability metrics (DuBay, 2004; Rabin et al., 1988).The existing quantitative approaches towards predicting readability of a text can be broadly classified into three categories (Benjamin, 2012):

Classical methods: they analyze the syntactic features of a text like sentence length, paragraph length etc. The examples are Flesch Reading Ease Score (Flesch, 1948), FOG index (Gunning, 1968), Fry graph (Fry, 1968), SMOG (McLaughlin, 1969) etc. The formulae do not take into account the background of the reader and the semantic features of the text such as whether the actual contents are making sense or not. Despite their shortcomings, these simple metrics are easy to calculate and provide a rough estimation of reading difficulty of a text provided.

Cognitively motivated methods: texts are analyzed based on the cognitive features like, cohesion, organization and users' background. Proposition and inference model (Kintsch and Van Dijk, 1978), prototype theory (Rosch, 1978), latent semantic analysis (Landauer et al., 1998), Coh-metrix (Graesser et al., 2004) are some prominent members of this group. This group of models moves beyond the surface features of a text and try to measure objectively the different cognitive indicators associated with text and the reader. However, it has been observed that, many situations, some traditional indicators perform as well as the newer and more difficult versions (Crossley et al., 2007).

Statistical language modeling: This class of approaches incorporates the power machine learning methods to the field of readability. They are particularly useful in determining readability of web texts (Collins-Thompson and Callan, 2005; Collins-Thompson and Callan, 2004; Si and Callan, 2003) (Liu et al., 2004). SVM has been used to identify grammatical patterns within a text and classification based on it (Schwarm and Ostendorf, 2005; Heilman et al., 2008; Petersen and Ostendorf, 2009). Although, these methods sound promising, the problem is that they cannot act as standalone measure:

they need an amount of training data for classifiers appropriate to a particular user group and often these measures takes into account complex text features which for resource poor languages need manual effort to annotate.

In Bangla, only a couple of works have been executed on text readability. Das and Roychoudhury (Das and Roychoudhury, 2006) studied a miniature model with respect to one parametric and two parametric fits. They have used seven paragraphs from seven literary texts. They considered two structural features of a text: average sentence length and number of syllables per 100 words. They found the two-parametric fit as better performer. Sinha et al. (Sinha et al., 2012) has developed two readability formulae for Bangla texts using regression analysis. For their study sixteen texts of length, about 100 words were used. They have considered six structural or syntactic features of a text for the work. They have demonstrated that the English readability formulae such as Flesch Reading Ease Index, SMOG Index do not perform appropriately while being applied to Bangla documents. They have found the textual features like average word length, number of polysyllabic words and number of jukta-akshars in a text to be the most influential ones. Both the works mentioned have taken into account a small subset of potentially important text features; none them have considered feature such as the extent of text cohesion. Moreover, their study did not explore the influence of readers' background on text readability. In our study, we have addressed the issue of readers' background as well as the effect of features at different textual level.

3 Empirical Data Collection

As mentioned, there is no annotated data present in Bangla, which can provide a direct classification of text difficulty for Bangla readers. Therefore, we have undertaken an effort to annotate the experiment texts with the target readers of Bangla.

3.1 Participants

Our objective in this study is to investigate how readability varies with the background of the reader. Therefore, two different target reader groups have been considered to study the relationship of effect of text parameters on comprehension and user background. SEC[1] or socio-economic classification has been stated according to the standards of Market Research Society of India (MRSI). MRSI has defined 12 socio-economic strata: A1 to E3, in the decreasing order. These strata have been designed based on the education level of the chief wage earner of the family and the number of "consumer durables" (as per a predefined list including agricultural land) owned by the family. It has been seen that this way of grading reflect the social and economic position of a household in terms of fields such as education, awareness etc. As can be inferred from the chart, the participants range from classes C2 to E1 (C2, D1, D2, E1), which represents the medium to low social-economic classes.

Type	**Background**	**Mean age (Standard deviation)**
Group 1 (adult): 25 native speakers of Bangla	Education: pursuing graduation	22.8 (1.74)
	SEC: C2-E1	
Group 2 (minors): 25 native speakers of Bangla	Education: pursuing secondary or higher secondary	15 (1.24)
	SEC: C2-E2	

Table1: User Statistics

3.2 Readability corpus preparation

We have stated in the introduction about the scarcity of annotated digital resource pool in Bangla useful for automatic processing. Although there are a few works on text readability in Bangla, the data is not available in accessible formats. To address the problem, we have developed a corpus of Bangla documents. The current size of the resource is about 250 documents of length about 2000 words spanning over broad categories such as News, literature, blogs, articles etc. A number of different text

[1] http://imrbint.com/research/The-New-SEC-system-3rdMay2011.pdf

features were computed against each document. The descriptions of the features and the justification for them have been stated below.

3.3 Feature selection:

Inferring from the cognitive load theory (Paas et al., 2003), we have assumed that the cognitive load exerted by a text on a reader depends on syntactic and lexical properties of a text like, average sentence length, average word length, number of polysyllabic words and as well as discourse features like the counts of the different parts of speeches and the number of co-references one has to resolve in order to comprehend the text. The logic behind such assumptions is as follows: while processing a text a user has to parse the sentences in it and extract semantically relevant meaning from those sentences and the words. In order to process a sentence, one has to take into account the length of the sentence and types of words contained in it; in addition, to infer the meaning of a sentence, it is important to establish the connections or the nature of dependencies among the different words in a sentence. The role of a word is determined by its parts of speech and its way of use in that context; apart from it, the words can have varied complexity based on factors like their length, count of syllables. Similarly, at the discourse level, a reader not only has to comprehend each sentence or paragraph, but also has to infer the necessary co-references among them to understand the message conveyed by the text. The complexity of this task depends on the number of entities (noun, proper nouns) in the text, how one entity is connected with other, relationships like synonymy, polysemy, and hyponymy. To capture the effects of all these parameters in our readability models, we have considered text features over a broad range. The details of the features are presented in Table 2. The word features like average word length, average syllable per word, sentence features like average sentence length and discourse features like number of polysyllabic words, number of jukta-akshars (consonant conjuncts) have been calculated as stated by Sinha et al. (Sinha et al., 2012), as the features need customizations for Bangla. The calculations based on lexical chains have been followed from Galley and McKeown (Galley and McKeown, 2003).

Feature	Description
word features	
average word length	Bangla orthographic word consists of a combination of four types of graphemes[2], each of them is considered as a single visual unit. Average word length is total word length in terms of visual units divided by number of words.
average syllable per word	Total word length in terms of syllable divided by total number of words.
sentence features	
average sentence length	Total sentence length in terms of words divided by number of sentence.
$(noun phrase)	Average number of NP per sentence
$(verb phrase)	Average number of VP per sentence
$(adjective)	Average number of adjectives per sentence
$(postposition)	Average number of postpositions per sentence. Bangla grammar has postpositions, instead of prepositions present in English. Unlike English, postpositions in Bangla do not belong to separate part of speech. The postpositions require their object noun to take possessive, objective or locative case. Suffixes act as the case markers.
$(entity)	average number of named entity per sentence
$(unique entity)	Average number of unique entity per sentence
$(clauses)	Average number of clauses per sentence

[2] http://en.wikipedia.org/wiki/Bengali_alphabet#Characteristics_of_the_orthographic_word

discourse features	
Number of polysyllabic words and normalized measure for 30 sentences	Polysyllabic words are the words whose count of syllable exceeds 2.
number of jukta-akshars (consonant conjuncts)	Total number of jukta-akshars in a text of 2000 words. It is an important feature for Bangla because each of the clusters has separate orthographic and phonemic (in some cases) representation than the constituents consonants.
#(noun phrase)	Total number of NP in the document
#(verb phrase)	Total number of VP in the document
#(adjective)	Total number of adjective in the document.
#(postposition)	Total number of postpositions in the document.
#(entity)	Total number of named entity in the document
#(unique entity)	Total number of unique entity in the document
#(lexical chain)*	Total number of lexical chain in the document
average lexical chain length*	Computed over the document

Table2: Details of text features considered for the study

The features marked with * in the above table have been manually annotated against each text. The other features, though they are computed automatically, a round of manual checking was incorporated for the sake of correctness.

Expert annotations and user annotations:

Since there is no formal ranking of Bangla texts according to their reading levels, therefore, the documents were then annotated by language experts to approximate the suitable reading level for each document. However, to develop any practical readability application, feedbacks from actual users are necessary. From the resource pool mentioned in Introduction, 135 texts were chosen for the present study: two sets of distinct 45 texts were for each group: for the adult group those were the texts annotated by experts to have relatively high reading level and for the minor's group, the texts were annotated as having relatively low reading level; pairwise t-test were performed between the two type of text features to assure that their difference is significant (*p*<0.05).

The rest 45 texts are common to both the groups to account for the difference in comprehension for the same document and the assumption that may in some cases group 2 participants have comparable reading skill as of group 1: consequently, the texts annotated by experts as demanding high reading level were selected for this purpose. These were required to ensure that the experimental data spans over a broad range and is unbiased. The text details are presented in table 2 below.

Source of Texts	Number of texts		
	Gr.1	Gr.2	common
Literary corpora_classical	5	5	5
Literary corpora_contemporay	6	5	6
News corpora_general news	6	6	5
News corpora_interview	5	6	6
Blog corpora_personal	6	5	5
Blog corpora_official	5	5	5
Article corpora_ scholar	6	7	7
Article corpora_general	6	6	6

Table3: Text details

Each participant was asked 2 questions: "How easy was it for you to understand/comprehend the text?" and "How interesting was the reading to you?". Against each question, they were to answer on a 5 point scale (1=easy, 5=very hard). Inter-rater reliability was measured through Krippendorff's alpha[3]

[3] http://en.wikipedia.org/wiki/Krippendorff's_alpha

and $\alpha = 0.81$ was found. Therefore, we concluded that annotators agree more often than would have occurred by chance. We have measured the correlation between the outcomes of two questions corresponding to each of the fifty annotators; and found that in each case the correlation was greater than 0.8 ($p < 0.05$). Therefore, the questions can be considered as equivalent, and subsequently we have considered the rating for the first question as user input for our readability models. Corresponding to each text, the average of the user ratings was considered for further processing.

4 Analysis and Model Development

4.1 Correlation coefficients

We have performed partial spearman correlation between each of the features and user rating. Table 4 presents some of the examples from each type of features due to the space limitation; results corresponding to other features are also described subsequently. The following features have selected as they have been used in the existing literature for Bangla (Sinha et al., 2012). The correlations are presented separately for the distinct texts and the common texts delivered to the two groups of users. This will allow us to investigate is there any significance difference of reading feedbacks between the different target populations.

Feature	Correlation coefficient r (Significance (if p<0.05) p value)			
	Different texts		Common texts	
	Gr. 1	Gr. 2	Gr.1	Gr. 2
Word features				
average sentence length	**0.8 (0.0017)**	**0.33(0.2011)**	**0.75 (0.0013)**	**0.54 (0.08)**
average word length	0.60 (0.0142)	0.73(0.0041)	0.66 (0.0026)	0.8 (0.0032)
Sentence features				
average syllable per word	**0.66 (0.06)**	**0.64(0.0047)**	**0.60(0.07)**	**0.75(0.0043)**
Discourse features				
number of polysyllabic words	0.73 (0.0013)	0.74 (0.0008)	0.67(0.0021)	0.65(0.0006)
normalized measure for 30 sentences	0.76(0.0011)	0.66 (0.0041)	0.65 (0.0015)	0.66(0.0032)
number of jukta-akshars	**0.87 (0.0018)**	**0.39 (0.1228)**	**0.81 (0.0024)**	**0.85 (0.0043)**

Table 4: Correlation coefficients (user rating vs text features)

Some interesting observations can be made from the above table:

- Average sentence length or mean number of words per sentence have been long found to be a strong predictor of text difficulty [1]. In our case, while this holds true for the adult data, the correlation is less for the minors and it is not significant.
- Average syllable per word does not hold significant correlation for the adult data in both cases but it does for the minor's group
- Jukta-akshars or consonant conjuncts have major impact on text readability in Bangla (Sinha et al., 2012). For adult data, it can be seen that this feature has a strong and significant correlation, which not true for the user data of group 2 for separate texts. On the other hand, for the common texts this feature was found to have high significant correlation with both the reader groups. This is may be due to the nature of the common texts.
- Apart from the above two cases, the above table also presents evidence in support of the fact that the reader's perception of text difficulty in relation to text features changes with the target reader background.

The impact of the remaining features has been discussed here with respect to the two different types of text scenarios:

Distinct texts for two groups:

- In case of the readers from the first group, the user ratings have high correlation ($r > 0.65$) with \$(clauses), #(verb phrases), #(unique entity), #(lexical chain) and average lexical chain length. The correlations are also significant. However, the correlations with \$(noun phrase), \$(verb phrase) \$(postpositions), #(postpositions), #(adjective) were found to be insignificant. The correlation of user annotation with features such as \$(entity), \$(unique entity) were found to be low ($r < 0.45$) but significant.
- The group 2 readers were found to show high ($r > 0.65$) and significant correlation with \$(verb phrases), \$(unique entity), \$(clauses), #(entity), #(lexical chain) and average lexical chain span. The correlations with \$(postposition), #(postpositions) were not significant. Features like \$(noun phrase), \$(adjective) and #(adjective) were found to have low ($r < 0.45$) but significant correlations with user ratings.

Common texts for both groups:

- It has been observed that the group 2 user ratings have higher correlation with the sentence level features than the discourse level features. In particular, features such as number of \$(noun phrase), \$(adjective), \$(unique entity) and \$(clauses) have high correlation with the text difficulty ratings provided by the minor's group. Among the discourse level features #(entity) and #(unique entity)have a high correlation, but #(verb phrase), #(adjective) were found to have not significant influence.
- On the other hand, the adult data are more inclined towards discourse features such as #(noun phrase) and #(verb phrase), #(unique entity) in a document. This may be due to the ability of the older people to comprehend the text as a whole rather than inferring meaning from individual units at a time. From sentence level feature \$(clause) was found to be significant and important in terms of correlation, but \$(noun phrase), \$(adjective) do not bear significant correlation.
- Properties like lexical chain, which require a reader to establish connections among different attributes of a concept have great significance for both group1 and group2 annotations.
- For both the user groups the influence of average \$(postposition and #(postposition) were found to be little and insignificant.

From the above discussions, it is evident that the two different target reader groups show a large difference in their reading pattern and perception of text difficulty. The difference has been observed in both the cases: when they were presented with different type of texts and with same texts. Therefore, it has been established that the target reader background plays an important role in modelling text difficulty. Accordingly, in the following sections, we have developed different models of different reader groups, and in the process we have also shown that the models have different parameter values and configurations.

4.2 Computational modelling

Analyses of correlation coefficients give an estimation of trend in user ratings against text features. The next step is to develop suitable models for automatic readability prediction. To achieve the objective, we have used machine-learning methods such as support vector machine (SVM) and support vector regression (SVR) techniques. In addition, we have also presented a comparative study of performances of different text features in readability model building in this section. The features have been used in three combinations. First they were divided in two categories i) comprising of only the six features mentioned in table 4 as they represent the 'classical' features used extensively to model text readability, and ii) second category consists of the rest 14 features and the group is termed 'non-classical' , this yielded the first two combinations. The third combination consists of all the features. Therefore, we have evaluated six different types of SVM and SVR models for each group.

We have employed a binary SVM classifier here. Given a training set instance-class pairs ($\overline{x}_i, y_i$), i = 1…l, where $\overline{x}_i \in R^n$ and $\overline{y} \in \{1, -1\}^l$, the general equation of a SVM is (Manning et al., 2008):

$$\frac{1}{2}\overline{w}^{T}\,\overline{w} + C\sum_{i}^{l}\xi_i \;\; is\ minimized,$$

$$\overline{w} = weight\ vector, C = regularization\ term \qquad \ldots(\text{equation: }1)$$

$$y_i\left(\overline{w}^{T}\,\Phi(\overline{x}_i) + b\right) \geq 1 - \xi_i, \qquad \xi_i(slack\ variable) \geq 0 \qquad \ldots\ (\text{equation: }2)$$

In this work, we have taken 90 texts against each group of users by combining the 45 reader group specific texts and 45 common texts (refer to section 3). Then for each category of reader, the texts were shuffled randomly. We have used 70 texts for training and 20 texts for evaluation of the model and performed 2-fold cross validation. The minimum, maximum and median of the rating distribution lie respectively at **(2.33)**, **(8.4)** and **(5.92)** for adult (group1) and at **(1.83)**, **(8.2)** and **(5.5)** for minor (group 2). To train and test the SVM models, we needed to spit the data in two classes (easy and hard), this has been done by assigning the ratings less than the median in to class easy (label '-1') and the rest to the class hard (label '1'), i.e., the user ratings were mapped to the label space $\overline{y}$. In case of SVR, the label space mapping was not required. The text features were mapped to the feature space $\overline{x}_i$. Although we have tested four types of kernel functions: linear, polynomial, radial basis and sigmoid on the data using LIBSVM (Chang and Lin, 2011) software, here only the results corresponding to linear and polynomial kernels have been presented as the other two kernels performed poorly. To evaluate the quality of the classifications for SVM, multiple correlation (R) and percentage of texts accurately classified (Acc) have been used. R denotes the extent to which the predictions are close to the actual classes and its square (R^2) indicates the percentage of dependent variable variation that can be explained by the model. Therefore, while percentage accuracy is an indicator to how well the model has performed to classify, R indicates the extent of explanatory power it posses. A better fit will have large R-value as well as Acc. For SVR, root mean square error (RMSE) has been reported instead of Acc; a good fit will have less RMSE. Below tables present, the SVM and SVR results for adult and minor's data for different kernels and different combination of features. The kernels were evaluated for a number of SVM parameter combinations and only the result corresponding to the most efficient one is presented.

Features	Classic features		Non-classic features		All features	
SVM parameters	C = 10; d = 2; r = 0; $\gamma = 1/6$ = 0.1; $\xi_i = 0.01$ (total support vector = 28)					
Kernel	R	Acc.	R	Acc.	R	Acc.
linear	0.75	76%	0.73	79%	0.80	87%
Polynomial	0.73	75%	0.72	75%	0.75	79.5%

Table 5: SVM for group1 readers

Features	Classic features		Non-classic features		All features	
SVM parameters	C = 1; d = 2; r = 0; $\gamma = 1/6$ = 0.1; $\xi_i = 0.001$ (total support vector = 22)					
Kernel	R	Acc.	R	Acc	R	Acc.
Linear	0.75	75%	0.72	77%	0.83	86%
Polynomial	0.71	70%	0.73	72%	0.78	76%

Table 6: SVM for group2 readers

Features	Classic		Non-classic features		All features	
Kernel	R	RMSE	R	RMSE	R	RMSE
linear	0.56	1.6	0.53	1.7	0.68	1.1
Polynomial	0.43	2.2	0.47	11.2	0.56	23.3

Table 7: SVR for group1 readers

Features	Classic		Non-classic features		All	
Kernel	R	RMSE	R	RMSE	R	RMSE
linear	0.50	1.5	0.54	1.4	0.65	1.2
Polynomial	0.47	3.1	0.45	15.5	0.51	29.7

Table 8: SVR for group2 readers

From table 5 and table 6, it can be seen that the SVM for the two target reader groups differ significantly in term of parameter attributes and their accuracy. It is also evident that incorporating only non-classic features versus classic features improves the accuracy of SVM very slightly and both types of features have similar explanatory power; combining both the classic and non -classic feature improves the accuracy and multiple correlations significantly. The SVR from table 7 and table 8 show the similar trend in terms of feature performances: classic and non-classis features have comparable RMSE and R, but there is significant gain when the two types are taken together. The regression equations for group1 and group2 readers differ in the coefficients of the feature variables; these imply that the two groups require different readability models. Moreover, the linear kernel was found to perform better than the polynomial kernel in all the cases.

5 Conclusion

In this paper, we have studied the effect of two important factors affecting text readability in Bangla: the target reader and text properties. We have found that the perception of text difficulty varies largely with the background of the reader. Accordingly, we have developed computational models to compute readability of Bangla text documents based on the target reader group. In order to achieve our goal we have first developed a novel Bangla dataset annotated in terms of text readability by users with varying age group. A preliminary analysis of the reading pattern of each target group was performed by analysing the correlation of text features with user annotations. Next, we have applied the SVM classifier to classify text documents into two different classes namely, *hard* and *easy*; the SVM for the two reader groups have different properties, implying the difference between two corresponding models. We have also compared the performance of the classifier based on the feature set they use. We observed that in contrast to applying only the classical features or the non-classic features, performance of the classifier improves if both types of features are used. This is true for both the adult as well as the minor’s dataset. Overall, we have achieved an accuracy of around 86% for the minor’s dataset and 87% for the adult dataset respectively. In addition to classification, support vector regression has been used to model text difficulty from an estimation perspective. The result of the SVR also establishes our previous findings. To the best of our knowledge, no such work on text readability has been recorded earlier in Indian languages, especially in Bangla. The next step of this study is to analyse the performance of the readability formula from one group (say adult) when applied to the other group (say minors) and vice versa. We will also repeat our study with more spread apart user groups spread over less diverse economic strata. In future, we are planning to develop for multi-class text readability models. The work will also be extended to model text comprehensibility for reading disabilities in Bangla.

Reference

Benjamin, R. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24:1–26.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Collins-Thompson, K. and Callan, J. (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL*, volume 4.

Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Dale, E. (1949). Readability.

Dale, E. and Chall, J. (1948). A formula for predicting readability. *Educational research bulletin*, pages 11–28.

Das, S. and Roychoudhury, R. (2006). Readability modelling and comparison of one and two parametric fit: A case study in bangla*. *Journal of Quantitative Linguistics*, 13(01):17–34.

DuBay, W. (2004). The principles of readability. *Impact Information*, pages 1–76.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Fry, E. (1968). A readability formula that saves time. *Journal of reading*, 11(7):513–578.

Galley, M. and McKeown, K. (2003). Improving word sense disambiguation in lexical chaining. In *IJCAI*, volume 3, pages 1486–1488.

Graesser, A., McNamara, D., Louwerse, M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202.

Gunning, R. (1968). *The technique of clear writing*. McGraw-Hill NewYork, NY.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79. Association for Computational Linguistics.

Kintsch, W. and Van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Landauer, T., Foltz, P., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Liu, X., Croft, W., Oh, P., and Hart, D. (2004). Automatic recognition of reading levels from user queries. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 548–549. ACM.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

McLaughlin, G. (1969). Smog grading: A new readability formula. *Journal of reading*, 12(8):639–646.

Paas, F., Renkl, A., and Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, 38(1):1–4.

Petersen, S. E. and Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106.

Rabin, A., Zakaluk, B., and Samuels, S. (1988). Determining difficulty levels of text written in languages other than english. *Readability: Its past, present & future. Newark DE: International Reading Association*, pages 46–76.

Rosch, E. (1978). Principles of categorization. *Fuzzy grammar: a reader*, pages 91–108.

Schwarm, S. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics.

Sherman, L. (1893). Analytics of literature: A manual for the objective study of english poetry and prose. *Boston: Ginn*.

Si, L. and Callan, J. (2003). A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems (TOIS)*, 21(4):457–491.

Sinha, M., Sharma, S., Dasgupta, T., and Basu, A. (2012). New readability measures for Bangla and Hindi texts. In *Proceedings of COLING 2012: Posters*, pages 1141–1150, Mumbai, India. The COLING 2012 Organizing Committee.