Educ Inf Technol
DOI 10.1007/s10639-014-9368-y

A study of readability of texts in Bangla through machine learning approaches

Manjira Sinha · Anupam Basu

Abstract In this work, we have investigated text readability in Bangla language. Text readability is an indicator of the suitability of a given document with respect to a target reader group. Therefore, text readability has huge impact on educational content preparation. The advances in the field of natural language processing have enabled the automatic identification of reading difficulty of texts and contributed in the design and development of suitable educational materials. In spite of the fact that, Bangla is one of the major languages in India and the official language of Bangladesh, the research of text readability in Bangla is still in its nascent stage. In this paper, we have presented computational models to determine the readability of Bangla text documents based on syntactic properties. Since Bangla is a digital resource poor language, therefore, we were required to develop a novel dataset suitable for automatic identification of text properties. Our initial experiments have shown that existing English readability metrics are inapplicable for Bangla. Accordingly, we have proceeded towards new models for analyzing text readability in Bangla. We have considered language specific syntactic features of Bangla text in this work. We have identified major structural contributors responsible for text comprehensibility and subsequently developed readability models for Bangla texts. We have used different machine-learning methods such as regression, support vector machines (SVM) and support vector regression (SVR) to achieve our aim. The performance of the individual models has been compared against one another. We have conducted detailed user survey for data preparation, identification of important structural parameters of texts and validation of our proposed models. The work posses further implications in the field of educational research and in matching text to readers.

Keywords Bangla text comprehensibility · Text readability · Resource creation · Readability models · Regression · Support vector machines · Support vector regression · User study

M. Sinha (✉) · A. Basu
Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India
e-mail: manjira87@gmail.com

Published online: 07 December 2014

1 Introduction

Reading is a complex cognitive action involving different steps like, recognizing and understanding individual words from a text, and decoding the grammatical structure of sentences to obtain the semantic information conveyed by them. Text readability or text comprehensibility generally refers to how well a reader is able to comprehend the content of a text through reading. According to Edgar Dale and Jean Chall (1948) readability is "*… the sum total* (*including all the interactions*) *of all those elements within a given piece of printed material that affect the success of a group of readers have with it. The success is the extent to which they understand it*, *read it at an optimal speed*, *and find it interesting*." We will use the terms readability and comprehensibility, interchangeably in the paper. For any given text, there is no absolute measure of difficulty based on purely the text parameters. The way a text will be interpreted depends on the background and context of the reader. Therefore, readability of a text has to be studied in relation with the characteristics of its reader, as well. Subsequently, we can assume that the cognitive load associated with the understanding of a text depends broadly on five factors. The first four are texts related parameters; they are often intertwined and the fifth one relates the reader:

1. **Lexical choice**: the complexity of the different words or phrases used in the text.
2. **Syntactic complexity**: the structural features of a text, it depends on the nature of sentences; their construction and inherent difficulties.
3. **Semantic complexity**: it represents the difficulty to grasp meaning from the words or sentences used in the text.
4. **Discourse level complexity**: it depends on text properties like cohesion, coherence, rhetorical structure of text.
5. **Background of the reader or the target audience**: it is a complex derivative of one's educational and socio-cultural background.

Apart from the reader and the text, the communicating language also affects readability. The structure and pattern of a language reflects how its users perceive and understand their surrounding world. Every language has some unique properties depending on its demography which in-turn control the determining factors for readability in that concerned language. Several cross-linguistic experiments indicate that language comprehension and processing are quite language dependent (Taft 2004). Therefore, the findings from experiments in one language cannot be generalized to all languages making it important to conduct similar experimentations in other languages. Consequently, different languages have developed different readability formulae (Rabin et al. 1988).

1.1 Importance of text readability

Readability of a text is a significant factor in the design of contents intended to match the reading competence of the target populations. Easy to read texts improve comprehension, retention, reading speed and reading persistence. The impact of readability research is felt whenever we need to have effective textual communication with people in different fields of activities, such as education, health-care, business or government policies. Expenditure overhead of most of the public welfare and awareness system increases largely due to lack of understanding of the information and instruction manuals. It has been found in studies that the average reading level of adult population in USA is 8th grade (Cotugna et al. 2005), which is why many public documents fail to meet their good intentions as they require higher comprehending abilities.

1.1.1 Importance of text readability in educational purposes

Research in text readability began from an educational perspective. It has been well established that the efficiency of educational contents for both children and adults increases if the comprehension difficulty and design of the text matches the target student group (DuBay 2007). Early researches such as Flesch Reading Ease Index, Dale-Chall readability formulae, Gunning Fog index etc. (refer to section 2) were dedicated to improving the school and college level reading materials. They focused on the reading difficulty of the educational materials and the actual reading ability of the target students groups. Subsequently, the scope of these researches extended beyond the conventional education sector to the areas like training guide for air-force and instruction manual for health education (DuBay 2004). Formulae like ATOS-TASA (Learning 2001) and Read-X (Miltsakaki and Troutt 2007) were developed by professionals in order to make school textbooks readable for students. Models such as proposition-inference by Kintsch and Van Dijk (1978) attempted to level text for matching the target reader population along with identifying the difficult areas. Methods like Latent Semantic Analysis (LSA) have been found to be effective for determining suitable educational material for college goers. More recent machine learning approaches by Heilman et al. (2008) and Petersen and Ostendorf (2009) have presented more in-depth analysis of educational contents, both in formal and informal sector. The studies mentioned above are only few examples from the numerous works on readability of educational materials. This brief overview suggests the importance of text readability in all levels of education.

Therefore we can conclude that readability has a two-fold importance in social aspects: first, in developing efficient contents for successful dissemination of education and literacy and second, in designing materials for successful conveyance of information to the target reader.

1.2 The context of Bangla

Literacy is the key to the socio-economic progress. In India, the adult literacy rate is well below the world average.[1] Low literacy rate implies lower reading levels, which in turn impedes empowerment of the common people. Moreover, India is a country with a large number of languages; according to census 2011 there are 1635 recognized mother tongue spanning over various language families and approximately 23 official languages that are used by different states.[2] Many of these languages have regional dialects. Although English is one of the official languages, a large section of Indian

[1] http://www.censusindia.gov.in/2011-prov-results/indiaatglance.html

[2] http://en.wikipedia.org/wiki/Languages_with_official_status_in_India#Eighth_Schedule_to_the_Constitution

population primarily use their mother tongue. Therefore, to reach to a large number of people, it is imperative to communicate in the native languages.

Despite the fact that readability measures are language dependent, the research on Indian language text readability is still in its infancy. In this paper, we have focused on identifying the readability of Bangla texts. Bangla is (also known as Bengali) is an eastern Indo-Aryan language and has its own script. The 220 million native and about 250 million total speakers have made it the seventh most spoken language in the world. Bangla is the second most spoken (after Hindi) and one of the official languages of India with about 84 million native users[3]; it is also the national language of Bangladesh. The need for focusing on one of the major native languages of India lies in the fact that people can interpret better, when the documents are in their own languages or mother tongue, i.e. their L1 language (Oakland and Lane 2004). However, even a native language instruction has to be comprehensible by the target reader; many welfare programs fail, as they require people to have a higher reading level than the present. Therefore, texts have to be designed, customized and presented in a manner that suits the cognitive capacity of target population. To meet this goal, at the very beginning, we have to identify how the different textual features affect the text difficulty in Bangla and how they can be modelled effectively.

1.2.1 Issue of usage variations and diglossia in Bangla

Bangla as used in India and Bangladesh posses some phonetic and accents (refer to footnote 3) and alternate lexical terms to denote some concepts. For example, to denote water, Bangla speakers in West Bengal, India mostly use *jola* (জল), while *pAni* (পানি) is used in Bangladesh, but these variations are primarily due to the religious influence on language rather than regional influence.[4]

Bangla has many dialectical variations in India as well as in Bangladesh (Agnihotri 2008). The standard form of Bangla used at present in both India and Bangladesh is based on the West-Central dialect of Nadia district (Agnihotri 2008). Bangla also exhibits diglossia i.e., the situation where two variants (or dialect) of the same language are used by a single language community (Ferguson 1959). In standard modern Bangla, the two diglossic variations are *sAdhubhAshA* (সাধুভাষা) or the highly codified version intended for formal documents and the more colloquial *calitabhAshA* (চলতিভাষা) form (Chakraborti 2003). The differences between the two forms are like the former uses more sanskritized vocabulary and longer verb inflections, where the later has comparatively simple grammar forms. However, in the present day usage of Bangla, apart from some legal and formal government documents, the *calita* form is used everywhere for both spoken and written communication.

In our study, we have focused on the *calita* (চলতি) form of the Bangla language as it is the only form used in real life purposes. Apart from some documents selected from the classic literature corpora (refer to section 4), all other test documents are in *calita* version. However, all the native Bengal users who participated in our experiment are accustomed with the *sAdhu* (সাধু) version as it is taught in school level language courses.

[3] http://en.wikipedia.org/wiki/Bengali_language

[4] We have used the convention: *Bangla term* (*iTrans transliteration*)

1.2.2 Difference between Bangla and english

In section 4.1, we have demonstrated that some of the widely used English readability formulae are not applicable for estimating text difficulty in Bangla. We concluded that this phenomenon is a result of the language structure difference between Bangla and English. Here, we have enumerated some of those points.

- Bangla and English belong to two very separate language families, namely Indo-European and West-Germanic.
- Bangla is morphologically richer than English. It has got a rich inflectional and derivational morphology inherited from Sanskrit, Persian and Arabic. It results in an abundance of compounding, and shows mild agglutination.
- Bangla orthography belongs to the *abugida*[5] class and English belongs to the *alphabetic* class. Therefore, the formation and identification of the visual form of a Bangla word happens in a different way than that of English.
- Almost all the readability metrics in English treat *polysyllabic* words (more than two syllables) as *hard* words, but in case of Bangla polysyllabic words are common in everyday use. For example, the verb *kariAchilAma* (করছিলাম) in Bengali corresponds to the phrase *I had done* in English and contains 5 syllables.
- Bangla is a *head-final* language that allows agreement with the rightmost conjunct when the verb follows the conjoined phrase.
- Bangla is also a *wh* in-situ language. In Bangla, the use of copular is not necessary for sentence construction. Both of these properties affect the skill required for sentence processing in Bangla.
- Bangla is a *flexible word order* language, which allows multiple grammatically correct surface forms.

Based on the above discussion, the objective of our work is to: a) create annotated readability resource in Bangla, b) development of computational models for text readability in Bangla and c) evaluation of the models. Our final aim is to design a input–output black-box system (see Fig. 1) which given an input text in Bangla will return some syntactic statistics for the text as well as its readability value as estimated by our readability formulae along with a result of binary classification (refer to section 5 for detail of model building).

At first, we have demonstrated that the well-known readability measures such as Flesch Reading Ease Index, SMOG index cannot be applied to determine text readability in Bangla. Then we have developed models in Bangla to predict overall reading difficulty of a text perceived by a native user. Our target user group (see section 4.2) consists of undergraduate or graduate level students belonging to medium to low economic background. We have used machine-learning methods such as regression, support vector machines and support vector regression to achieve our aim. The outcomes of the models are discussed in context of text comprehensibility and are compared against each other. Our

[5] http://en.wikipedia.org/wiki/Abugida

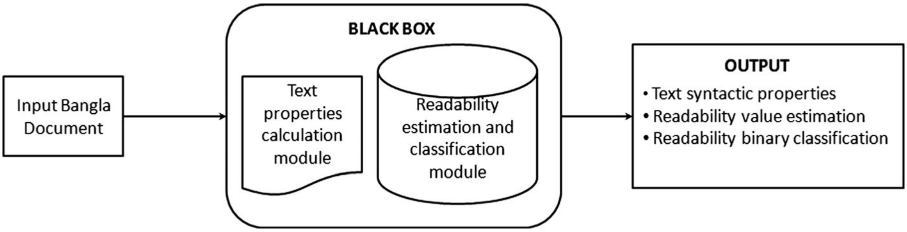

Input Bangla Document
BLACK BOX
Text properties calculation module
Readability estimation and classification module
OUTPUT
• Text syntactic properties
• Readability value estimation
• Readability binary classification

Fig. 1 Outline of a readability calculating system

study is based on the structural or syntactic features of a text like average sentence length, average word length in terms of visual units and we have customized definitions of these features to accommodate the specificities of Bangla (refer to the "Data Preparation" section).

The organization of the rest of the paper is as follows: section 2 presents a brief literature survey of text readability; section 3 describes text preparation and feature selection; section 4 illustrates the inapplicability of English readability metrics for Bangla and the subsequent user study that has been undertaken. Section 5 details the computational procedures and the inferences we have drawn; this section is followed by the final section on conclusion and future works.

1.2.3 Related works

The objective of readability studies was to develop practical and easy to implement methods to grade texts according to the reading abilities of adults (Buswell 1937). The quantitative analysis of text readability started with L.A. Sherman in 1880 (Sherman 1893). Until date, English has over 200 readability metrics. Studying readability is gaining much popularity (DuBay 2004). Attempts have also been made in other languages such as Spanish, French, German, Dutch, Swedish, Russian, Hebrew, Chinese, Vietnamese and Korean (Rabin et al. 1988). The existing quantitative approaches towards predicting readability of a text can be broadly classified into three categories (Benjamin 2012) –these categories are not distinct, many a times they overlap and communicate with each other:

1.3 Classical methods

This type of approaches incorporate 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 chronologically newer formulas like new Dale-Chall index (Chall 1995), lexile framework (Stenner 1996), ATOS-TASA (Learning 2001), Read-X (Miltsakaki and Troutt 2007) consider the readers' background and text semantics by incorporating information like word familiarity, word frequency or graded vocabulary list etc. We have considered the following four models to examine their applicability in Bangla; the reason is their high correlation with the established comprehension tests in English (DuBay 2007; McLaughlin 1969):

11 **Flesch Reading Ease**=206.835−(1.015×average sentence length)−(84.6×average number of syllables per word)

23 **Flesch-Kincaid Grade-Level**=(0.39× average sentence length)+(11.8×average number of syllables per word)—15.59

32 **Gunning FOG grade**=0.4 (average sentence length+percentage of Hard Words)

44 **SMOG grading**=3+square root of polysyllable count per 30 sentences

The formulae from this class fail short in some major situations. They do not take into account the background of the reader and only measure the surface level features of a text. They do not consider the semantic features of the text such as whether the actual contents are making sense or not. Despite their shortcomings, these simple metrics are still useful for many purposes. They are easy to calculate and provide a rough estimation of reading difficulty of a text provided.

1.4 Cognitively motivated methods

This class of methods uses high-level text parameters like cohesion, organization and cognitive aspects of the reader. Proposition and inference model (Kintsch and Van Dijk 1978), prototype theory (Rosch 1978), latent semantic analysis (Landauer et al. 1998), semantic networks (Foltz et al. 1998) are examples of this category. This type of approach introduced text levelling or text revising methods (Kemper 1983; Britton and Gülgöz 1991). One distinguished instance of this class is Coh-metrix (Graesser et al. 2004), (Graesser et al. 2011)], which scores a text based on 200 different features spanning over a broad range including sophisticated features like text cohesion. Coh-Metrix has been used to measure text difficulty for both L1 and L2 (Crossley et al. 2007). Recently Coh-Metrix is being used to assess writing quality and to distinguish between relative cohesion of texts (McNamara et al. 2010). The DeLite software (vor der Brück et al. 2008) uses a dedicated syntactic-semantic parser to analyze a text in German. It predicts the hardness of a text based on textual parameter values from five different levels: morphological, lexical, syntactic, and semantic and discourse. It also incorporates machine-learning algorithms to normalize the parameter values and determine indicator weights. This in-turn improves its performance, DeLite's prediction has been found to be more correlated with user predictions than that by traditional formulas (Benjamin 2012). DeLite serves as a bridge between cognitive motivated methods and statistical language modelling techniques.

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, such studies are still in infancy. One of the major drawbacks of these techniques is that they can be too complex to be implemented for practical purposes. They require automation to an extent. Moreover, in has been observed that, in many situations, some traditional indicators perform as well as the newer and more difficult versions (Crossley et al. 2007).

1.5 Methods involving statistical language modelling

This class of approaches incorporates the power of machine learning methods to the field of readability. They expand the traditional simple indices to a probabilistic analysis. They are particularly useful in determining online readability based on user queries (Liu et al. 2004). They have been used to predict readability of web texts (Collins-Thompson and Callan 2005; Collins-Thompson and Callan 2004; Si and Callan 2003). Sophisticated machine learning methods like support vector machines (SVM) have been used to identify grammatical patterns within a text and classification based on it for both web texts (Heilman et al. 2008) and traditional texts (Schwarm and Ostendorf 2005; Petersen and Ostendorf 2009). Schwarm and Ostendorf (2005) have also used support vector machines for error analysis in text difficulty with respect to the intended grade level.

Although, these methods sound promising, the shortcoming is that they cannot act as standalone measures as they need an amount of training data for classifiers appropriate to a particular user group. Moreover, they also need extensive user study for training and validation.

1.6 Work done in Bangla

Compared to numerous readability measures in English (Benjamin 2012), few initiatives have been taken in Bangla. Das and Roychoudhury (2006) studied a miniature model with respect to one parametric and two parametric fits. They considered two structural features of a text: average sentence length and number of syllables per 100 words. Seven paragraphs for seven different texts were used. They found the two-parametric fit as better performer. Islam et al. (Islam et al. 2012) have performed readability classification on 24 textbooks in Bangla from the *National Curriculum and Textbook Board*, *Bangladesh*.[6] The books were ranging from class two to class eight and the reading difficulty of a text was assumed based on the class level only. They have developed a readability classifier based on 25 information-theoretic features and five lexical features. They have found that the combination of all the features yield an accuracy score of 75 %. However, their work do not describe the classification models or target classes, moreover, as the domain of texts are restricted to only school textbooks, the performance of the features over general Bangla documents are yet to be evaluated.

1.6.1 Data preparation

Identifying the structural parameters of a text We have already discussed that Bangla has many distinct characteristics from English. Bangla has a set of 53 characters including 13 vowels and 40 consonants and has around 198 distinct consonant clusters with individual graphemic forms. We have considered the following standard structural or syntactic parameters of a text but customized them to accommodate the specificities of Bangla:

[6] http://www.nctb.gov.bd/

1. **Average Sentence Length** (**ASL**): Sentence length accounts for the number of words separated by spaces or any symbol in a sentence. Sentences are separated by *purnaviram* or a dividing punctuation mark (a question mark or an exclamation symbol). A *purnaviram* is equivalent to a full stop/period in English to mark the end of the matter. Average Sentence Length is computed as dividing total sentence length by total number of sentences.
2. **Average Word Length** (**AWL**) **in terms of visual units**: Along with dedicated graphemes for consonants and vowels, Bangla scripts have some additional graphemes corresponding to the vowel modifiers (diacritic) and consonant conjuncts (jukta-akshars). Bangla orthographic word, is combination of the following kinds of graphemes[7]:

a. Independent form of vowel graphemes.

b. Consonant graphemes with or without a vowel diacritic attached to them.

c. Consonant conjuncts with or without a vowel diacritic attached to them.

d. Other modifier symbols indicating nasalization of vowels, and suppression of the inherent vowels.

We consider each kind as a separate visual unit of a word, which is equivalent to each alphabet in an English word. The length of a word corresponds to total number of visual units in that word. Average Word Length is equal to total word length divided by number of words. An example is given below:

(**dibAniShi**)=দি(**di**) + বা (**bA**)+ নি(**ni**)+ শি(**Shi**) Length=4, all the units here represents the second type (b) of graphemes specified above.

3. **Average number of Syllables per Word** (**ASW**): A syllable is a unit of organization for a sequence of speech sounds. For syllable count in a Bangla word, we use Bangla grapheme to phoneme converter tool (G2P).[8] Average number of Syllable per word is equal to total syllable count divided by number of words.
4. **Number of PolySyllabic Words** (**PSW**): Polysyllabic words are the words whose count of syllable exceeds two.
5. **Number of PolySyllabic Words per 30 sentences** (**PSW30**): Polysyllabic words per 30 sentences are computed by taking the number of polysyllabic words in total text, dividing it by total number of sentences and then multiplying it by 30.
6. **Number of Jukta-akshars** (**JUK**) **or consonant clusters**: jukta-akshar or consonant-conjunct is the circumstance when consonants or vyanjan occur together in clusters. When a consonant with a halant (hasanta) is followed by another consonant, we consider it as one jukta-akshar. More than one consonant with halants, followed by a full consonant, is also considered as one jukta-akshar. A consonant occurring at the end of a word, i.e. it is not followed by any other consonant, is not considered as a jukta-akshar. The number of jukta-akshars count is the total number jukta-akshars present in the text. Jukta-akshars are an important feature for Bangla because each of the clusters has separate orthographic and phonemic (in some cases) representation than the constituents consonants. The measure is normalized for 50 sentences so that it can be compared across different texts.

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

[8] Downloaded from www.cel.iitkgp.ernet.in

An example of a jukta-akshar is ক্ষ(ksha)=ক (ka)+্(halant)+ষ (sha)

Therefore, শিক্ষা (shikshA)= ি (i)+শ (Sha)+ক(ka)±্(halant)±ষ(sha)+ া (A) has jukta-akshar count equals to one.

The last one, i.e. the number of jukta-akshars (JUK) is an important structural parameter for Indian language text. Although English has consonant conjuncts like *pt* in some exceptional cases, the role hold by jukta-akshars in Bangla are more widespread and significant. The relation between jukta-akshars and text readability is being examined for the first time.

1.6.2 Text selection

In Bangla We could not find any publicly available Bangla dataset annotated in terms of their reading difficulties. As a result of this, 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. The current size of the resource is about 200 documents of length 2000 words spanning over broad categories such as News, literature. For our study, we have randomly selected seventy-five (75) texts. The details are provided in Table 1 below:

1.7 Applying existing english readability measures on Bangla texts

At first, Bangla texts are analyzed with the established readability indicators for English. The scores obtained by the English models are presented in Table 2 below (for the sake of convenience, only 11 texts are presented). The table depicts the results from four famous and mostly trusted (Chall 1958; Klare 1963) English readability models namely: Flesch Reading Ease score (ER 1), Flesch-Kincaid Grade (ER 2), Gunning Fog index (ER 3) and SMOG index (ER 4). Although these readability Models have been applied to several languages with satisfactory results (Bamberger and Rabin 1984), in our case out of bound results are found. As an instance, reading score of Flesch Reading Ease should lie in the range of 0–100 with a highest possible upper bound of 120 (Flesch 1948), whereas for Bangla texts, its value is more than 150

Table 1 Details of texts with category, quantity and size

Source of texts	Number of texts	Words (approximated in thousand (k))
Literary corpora_classical	9	19
Literary corpora_contemporary	10	20
News corpora_general news	10	15
News corpora_interview	9	15
Blog corpora_personal	9	19
Blog corpora_official	9	20
Article corpora_ scholar	10	22
Article corpora_general	9	21

for all the texts. Results of Flesch-Kincaid Grade Level are not even positive and smaller than −4, whereas, the possible lowest grade in theory is −3.40 with few real situations[9] (Kincaid et al. 1975). Grade levels evaluated by Gunning Fog Index and SMOG Index are within their prescribed range, but validating those with the actual reading standards of the experiment texts reveal that they are far from the expected grades[10] (Gunning 1968; McLaughlin 1969). For example a Gunning Fog Index score of 18 requires an educational level equal to post graduation or more, but as per our analysis, almost all of the experimental texts yield an index more than 18 including the texts from standard 8 school text books.

From the above results, we can infer that these models are not suitable for measuring the readability of Bangla texts. The disagreement on the values can be attributed to the significant differences in the language structure of English and Bangla as discussed In our earlier sections.

Therefore, the above observations motivated us to develop a new model of text readability specifically for the Bangla language. Accordingly, we have performed user studies to collect readability judgments on a large number of Bangla texts from different group of users. These collected dataset is used to develop computational models of the reading difficulty of Bangla texts. The user survey has been described in the following section.

1.8 User study

1.8.1 Participants

Choice of target reader group is a significant part of study on text readability due to the highly subjective nature of text difficulty. Fifty native speakers of Bangla participated in the user study. To reflect reading difficulty experienced by average Bangla native population, we have selected the participants based on two criteria: Socio-Economic Classification (SEC) and proficiency in Bangla language. The socio-economic condition and the educational backgrounds of the participants are provided in Fig. 2a and b below, in the form of pie charts. The socio-economic classification was obtained according to the guidelines provided by The Market Research Society of India (MRSI) in their SEC classification manual.[11] MRSI has defined 12 socio-economic strata: A1 to E3, in the decreasing order. As can be inferred from the chart, the participants range from classes B1 to C2, which represents the average or medium social-economic classes. Moreover, a correspondence with the mean Household Potential Index (HPI) shows distribution from 18.7 to 4.4, which also represents the medium range of HPI. To capture the language skill, each native speaker was asked to rate his/her proficiency in Bangla on a 1–5 scale (1: very poor and 5: very strong) as can be seen from Fig. 2c, majority of the selected participants have proficiency medium to poor. In the backdrop of a country like India it is not exceptional that a person pursuing graduation or higher education is from a

[9] http://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests

[10] http://en.wikipedia.org/wiki/Gunning_fog_index

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

Table 2 English readability models applied to Bangla

Text No	ER 1	ER 2	ER 3	ER 4
1	170	−4	24	19
2	172	−5	23	18
3	173	−5	21	17
4	183	−9	13	11
5	173	−5	21	17
6	178	−7	16	13
7	174	−6	20	15
8	177	−7	20	15
9	161	−1	30	23
10	168	−4	25	19

medium to low economic background and is not so proficient in the native language. The choice of participants representing average population partially met our long-term goal to model reading difficulty in Bangla of the backward social-economic classes. The objective we will achieve in our subsequent readability studies in future.

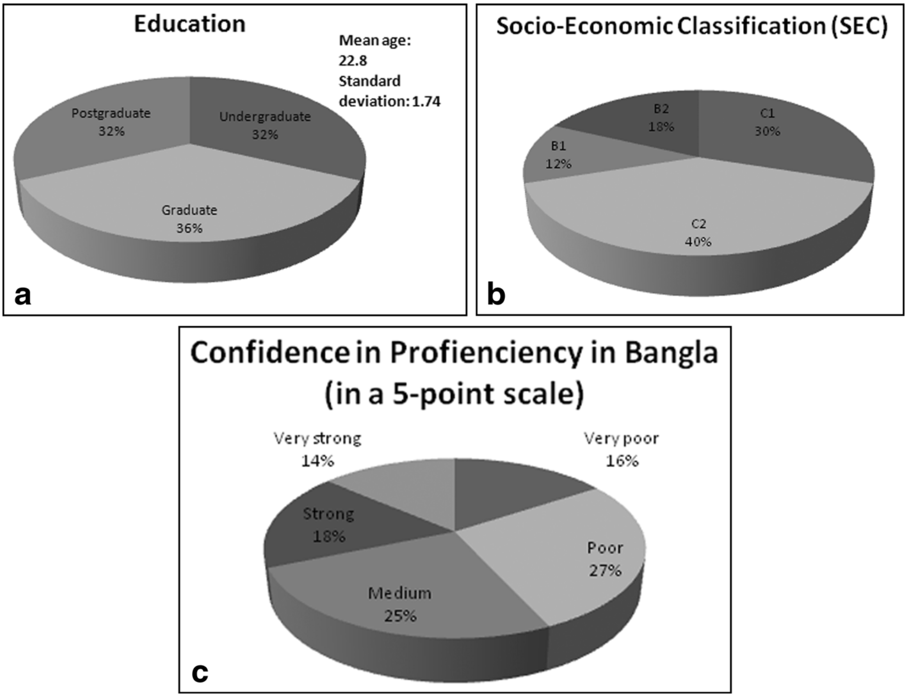


Fig. 2 Participants' details: **a** (*top-left*) education and age, **b** (*top-right*) social and economic background, **c** (*bottom*) confidence in proficiency in Bangla

1.8.2 Procedure

Each participant was presented with the 50 texts mentioned above. They were instructed to read the text carefully. Upon completion, they were asked the question: "how easy was it for you to understand/comprehend the text?", and they were to answer on a scale of 10, where '1'stands for very easy to understand/comprehend and '10' for extremely difficult. The averages of all users' score against the 75 documents are presented in form of a histogram (Fig. 3). It can be seen that the documents were chosen over a broad range of spectrum to incorporate different levels of reading difficulty. The user ratings were validated with 2σ validation around the mean with a 99 % confidence interval.

As a simple way to check the degree of variation of different linguistic features to the evaluation done by the users, the Spearman's rank correlation (Zar 1998) has been computed between them. Figure 4 presents correlation between the features and the user ratings. From Fig. 4 above, it is visible that the factor that mostly correlates with the user's perception of hardness of a text is the number of jukta-akshars present per 50 sentences in the text. The next are the number of polysyllabic words followed by average word length in terms of visual units and then the number of average syllable per word. Interestingly, the average sentence length, although found to be an influential factor in text readability (Crossley et al. 2007) comes at only fourth in terms of the correlation. After analyzing the correlation of textual parameters with user perception, we next move onto computational model building. One point must be noted here that, although attempts are being made to develop readability measures that are language independent, they are comprised of two parameters: average sentence length and average word length (Grzybek 2010). Therefore, we need to observe whether at all these two parameters in their conventional meanings are significant in Bangla text readability.

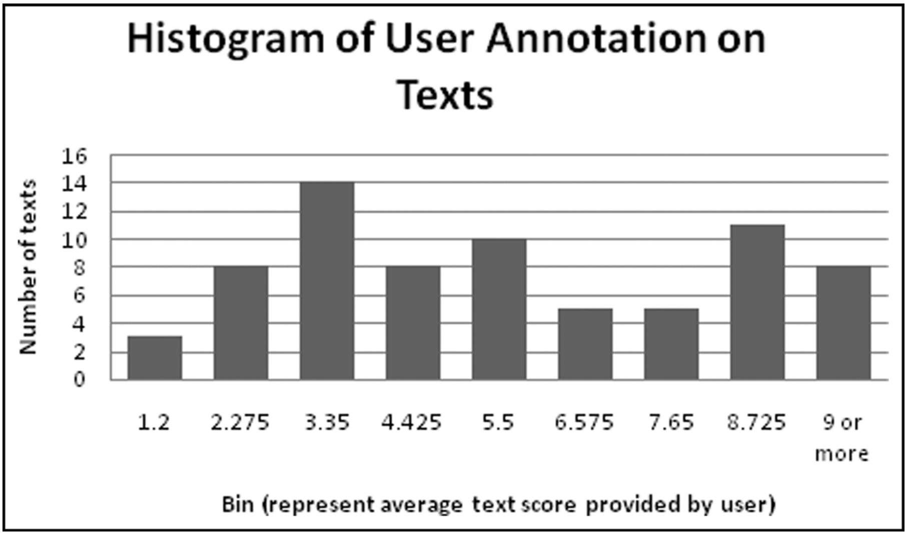


Fig. 3 Distribution of user annotations

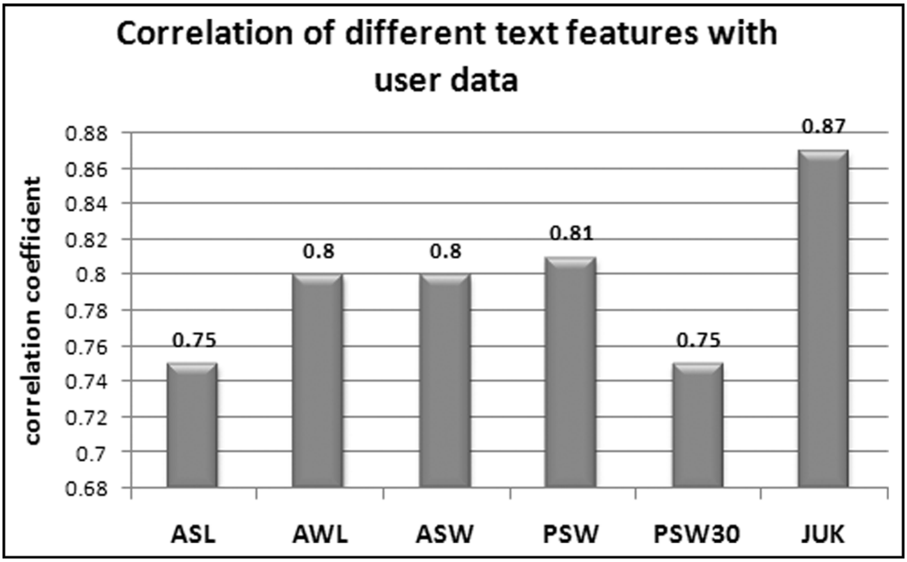


Fig. 4 Correlation of text parameters with user scores

1.9 Computational models for text readability prediction

Differentiating or grading texts according to their reading level can be viewed as a classification problem as well as a regression problem. We have analyzed both of the approaches in the following subsections. The classification approach differs from the regression approach in certain ways as they have different objectives. In the following sections, we have elaborated the statistical and machine learning techniques as well as the implications of the results obtained by their application to our problem. We felt the explanations of the background objectives of the techniques are necessary as these objectives will interpret the results from the corresponding methods.

Regression analysis (Montgomery et al. 2007) is an *estimation* problem as it predicts the absolute value of the dependent variable based on the independent terms; it attempts to minimize the *sum of squared error* (SSE).[12] The goodness of regression can be judged by *coefficient of determination* (R^2)[13] Most of the traditional readability indices mentioned in section 2 have been developed using regressions. On the other hand, Support Vector Machine or SVM (Cortes and Vapnik 1995; Manning et al. 2008) is a *large-margin classifier* that attempts to partition the test data in two distinct groups with as much distance as possible, based on an earlier training set. SVM can be beneficial compared to regression due to the following reasons: SVM (especially, soft-margin classifier) has a strong theoretical basis to *regularization*, which prevents the classifier from *overfitting* the data; it also performs well when there are many attributes and few cases to train the model. As discussed above in related works, SVM method is often found to be more effective than regression in predictive difficulty level of a text (Petersen and Ostendorf 2009).

[12] http://en.wikipedia.org/wiki/Mean_squared_error
[13] http://en.wikipedia.org/wiki/Coefficient_of_determination

When applied to text difficulty analysis, regression technique assigns an absolute *hardness score* to the target document, whereas, SVM classification determines if the target document belongs to the hard or easy class. Therefore, whether to use either of the two techniques depends on the intended use.

1.9.1 Model development by regression

Training For the training of the regression models, we have used 50 out of the 75 texts. We have examined the possible one-parametric (linear and hyperbolic), two-parametric (linear) and three-parametric (linear) fits. Table 3 documents the short-listed models (including Flesch and SMOG equivalence (model 1 and model 2)) for which the fitting is optimal (low SSE and high R^2) from each category. Although, we have calculated the one-parametric hyperbolic fits on each text parameters, their R^2 turn out to be negative suggesting that this kind of model is not fit for our cases.

Validation We have applied our six short listed readability models (Table 4) to the remaining 25 texts of the 75 texts for validation purpose. The root mean square error (RMSE) has been taken as a measure of the accuracy of the models. The RMSE of the predictions made by our readability models compared to the actual scores given by the users are summarized below in Table 4. Figure 5 represents the comparison of performances of the six different models. At the training phrase, model 3, model 4, model 5 and model 6 have comparable results, but during validation, clearly, model 3 and model 4 have significant better results over the other two models (refer to Table 4). In addition, it is to be noticed that these models have number of jukta-akshars per 50 sentences (JUK), average word length in terms of visual units (AWL) and number of polysyllabic words (PSW) as the parameters with positive coefficients and number of polysyllabic words per 30 (PSW30) sentences with a negative coefficient. Moreover, in model 3, PSW has very small coefficient implying they have comparatively less contribution in text difficulty. The equivalence of Flesch Reading Ease (model 1) and SMOG index (model 2) which rely on ASL, ASW and sqrt (PSW30) respectively are not in the top, though these two indices have significant impact for English. Therefore, it is established that the efficient readability models of Bangla do not have ASL and ASW as their two parameters. Subsequently Bangla text readability cannot be considered along with the research towards development of language independent readability

Table 3 Tentative readability metrics in Bangla

Model	Expression	R^2	*SSE*
Model 1	−10.4+.11*ASL+5.22*ASW	0.58	1.77
Model 2	0.44*sqrt (PSW30) −1.79	0.53	N/A
Model 3	−5.23+1.43*AWL+.01*PSW	0.80	0.82
Model 4	1.15+.02*JUK−.01*PSW30	0.78	0.9
Model 5	5.37+.01*PSW−2.29*ASW+.01*JUK	0.83	0.83
Model 6	5.71+.18*ASL−1.49*ASW+.01*PSW	0.83	0.84

Table 4 Summary of validation results

Model	1	2	3	4	5	6
RMSE	1.32	1.19	0.54	0.67	1.08	1.27

models, which rely on the parameters ASL and ASW (Grzybek 2010); instead, we need to have novel readability models for Bangla. Accordingly, we propose these two models (model 3 and model 4) as our readability metric for Bangla, namely: ReadabilityinBangla1 (RB1) and ReadabilityinBangla2 (RB2). Fig. 6 presents a visualization of the predictions obtained by the two metrics along with the user feedback for the 25 validation texts Figs. 7 and 8.

1.9.2 Classification using support vector machines (SVM)

In this section, we have applied the binary SVM classifier to classify Bangla text documents into two binary classes namely, *hard* and *easy* texts. Since, the median value of the user ratings is 4.9, we have labeled the texts having user rating less than 4.9 as easy or class -1 and the rests as hard or class 1. Therefore, the feature space $\overline{x}(\overline{x}\epsilon R^n)$ of our SVM consists of a 75×6 matrix containing 6 features (mentioned in Data Preparation section) of each of the 75 texts and the binary mapping of user evaluation corresponding to the texts represents the label space. Given a training set instance-class pairs $(\overline{x}_i,\ y_i\ )$, $i=1\ldots l$, where $\overline{x}_i\in R^n$ and $\overline{y}\in\{1,\ -1\}^1$, the general equation of a SVM is (Manning et al. 2008):

$$\begin{gathered}\frac{1}{2}\overline{w}^T\,\overline{w}+\ C\sum\nolimits_i\xi_i\ \mathit{is\ minimized},\\ \overline{w}=\mathit{weight\ vector},C=\mathit{regularization\ term}\end{gathered}\tag{1}$$

Table 5 Results of SVM classification on Bangla text using different kernels

	Linear	Polynomial	Radial basis	Sigmoid
Fraction of texts correctly classified	80 %	75 %	56.25 %	56.25 %
multiple correlation (R)	.87	.85	.67	.67
C=10;d=2;r=0;γ=1/6=0.1;ξ_i=0.01				
Fraction of texts correctly classified	75 %	73 %	56.25 %	56.25 %
multiple correlation (R)	.81	.79	.65	.65
C=10;d=2;r=0;γ=1/6=0.1;ξ_i=0.001				
Fraction of texts correctly classified	70 %	70 %	55 %	55 %
multiple correlation (R)	.79	.72	.63	.64
C=100;d=2;r=0;γ=1/6=0.1;ξ_i=0.01				

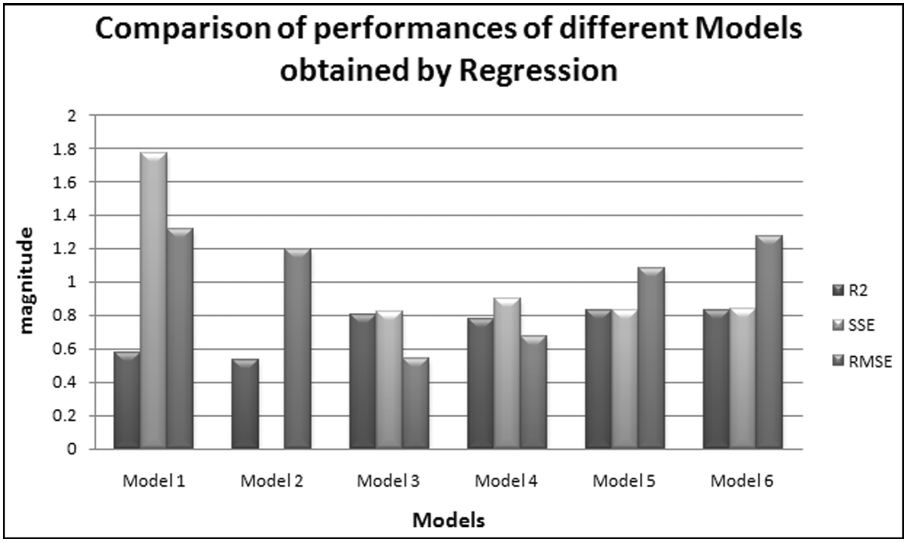


Fig. 5 Visualization of performance of different regression models

$$y_i\left(\overline{w}^T \Phi(\overline{x}_i) + b\right) \geq 1 - \xi_i, \quad \xi_i(slack\ variable) \geq 0 \tag{2}$$

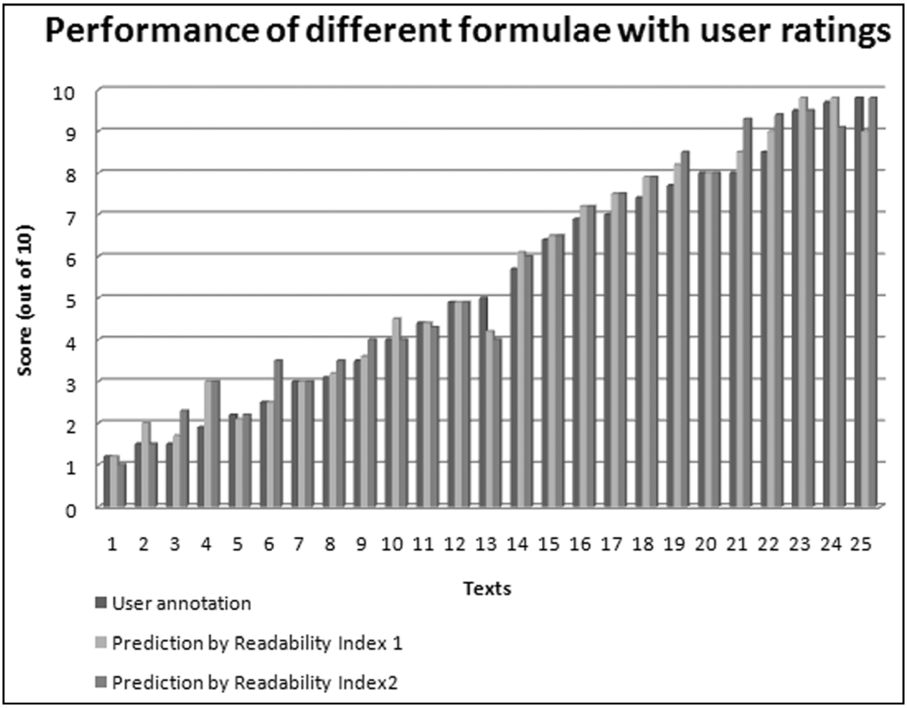


Fig. 6 Comparison of model performances with user ratings

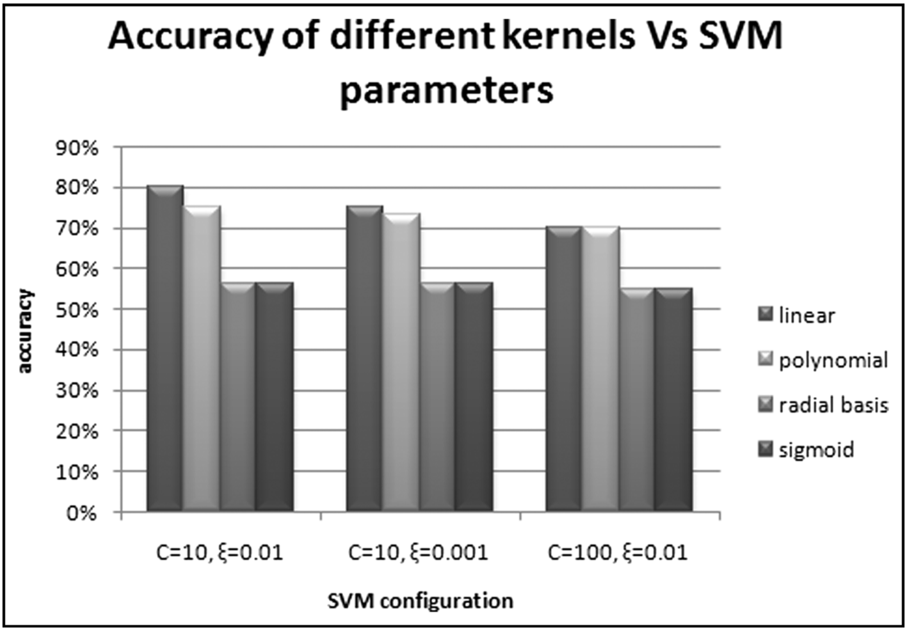


Fig. 7 Accuracy of different kernels with respect to the SVM parameters

We have divided the dataset in 4 parts, each having 4 texts and have performed a 4-fold cross validation. We have used four types of kernel function on the data using LIBSVM (Chang and Lin 2011) software:

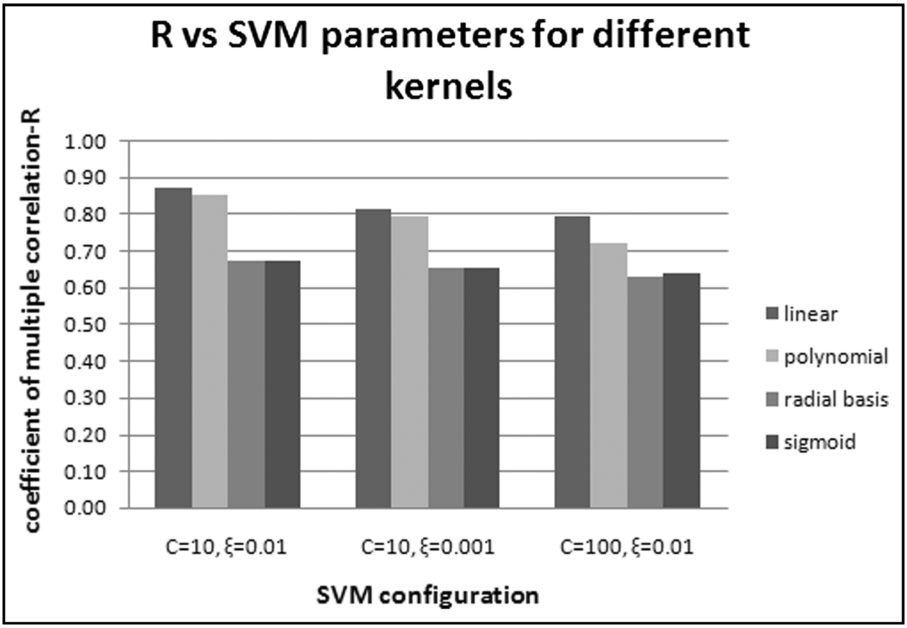


Fig. 8 Comparing coefficient of multiple correlation-R and SVM parameters for different kernels

1. linear : $K\left(\overline{x}_i, \overline{x}_j\right) = \overline{x}^T{}_i\overline{x}_j$
2. polynomial : $K\left(\overline{x}_i, \overline{x}_j\right) = \left(\left(\gamma\overline{x}^T{}_i\overline{x}_j\right) + r\right)^d, d = \text{degree}, \gamma = 1/features$
3. radial basis function : $K\left(\overline{x}_i, \overline{x}_j\right) = \exp\left(-\gamma\left(mod\left(\overline{x}_i - \overline{x}_j\right)\ ^2\right)\right)$
4. sigmoid : $K\left(\overline{x}_i, \overline{x}_j\right) = \tanh\left(\left(\gamma\overline{x}^T{}_i\overline{x}_j\right) + r\right)$, $r =$ coefficient

The table above presents the results of SVM classifications with different set of SVM parameters as mentioned in Eqs (1) and (2) along with different types of kernels. To evaluate the quality of the classifications, multiple correlation (R) and percentage of texts accurately classified are used. R denotes the extent to which the predictions are closed to the actual classes and its square (R^2) indicates the percentage of dependent variable variation that can be explained by the model. As can be shown from the results above, about three-fourth of the documents were classified correctly using linear or polynomial kernel of degree 2, when the regularization term C=10 and the value of the slack variable $\xi_i \leq 0.01$. Therefore, we can conclude that *hard* and *easy* to comprehend texts are linearly classifiable in the 6-dimensional document feature space mentioned in section 3. Compared to the regression approach, in case of SVM, it was not necessary to choose a subset of the text features in order to obtain good results.

1.10 Support vector regression (SVR)

Support vector regression (SVR) extends the SVM technique into estimation problems such as regression (Drucker et al. 1997; Basak et al. 2007; Smola and Schölkopf 2004). We have examined the performance of SVR in text readability determination. Linear and polynomial kernel of degree 2 has been used along with the epsilon loss function value at 0.1. We have used two different combinations of features: first, all the six features were considered and then the three features (AWL, PSW, and JUK), which were found to be the most influential from the results of linear least square regression (refer to section "model development by regression") were taken. The results are presented in the Table 6. As can be deduced from the analysis of R^2 and RMSE values, SVR performs poorly than the least square regression when applied to text difficulty analysis (refer to Tables 4 and 5). Fig. 9 presents the comparative chart of RMSE values as obtained by least square regression and SVR.

As regression and classification approach address the problem of text readability in different ways, it is difficult to compare their outputs in absolute terms. But for the sake of convenience, we have presented a comparison of the Goodness of Fit ($R^{2)}$ for linear regression, SVM binary classification, SVM regression with 6 features, SVM regression with three features (refer to Fig. 10). The two series corresponding to linear regression represents the two readability formulae developed, and the series

Table 6 Results of SVR

	Linear kernel	Polynomial kernel		
	RMSE	R^2	RMSE	R^2
All features	1.6	0.02	2.2	0.63
Three features	1.1	0.28	23.3	0.47

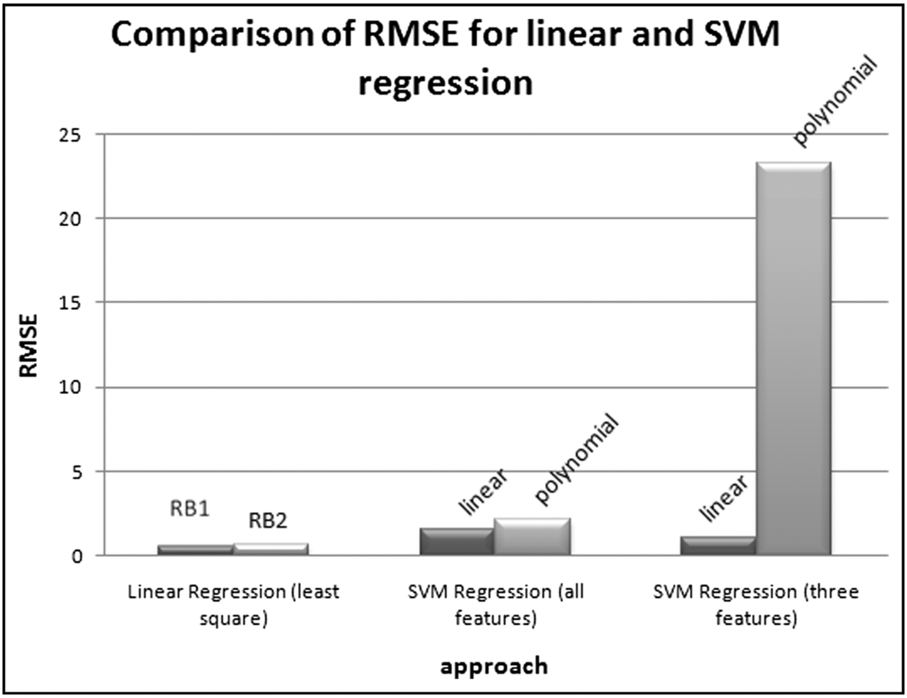


Fig. 9 Comparison of RMSE for linear and SVM regression

corresponding to SVM and SVR denotes performances of linear and polynomial kernels. As can be seen from the charts, both linear regression and binary SVM classifier have performed efficiently in achieving their desired objectives and have

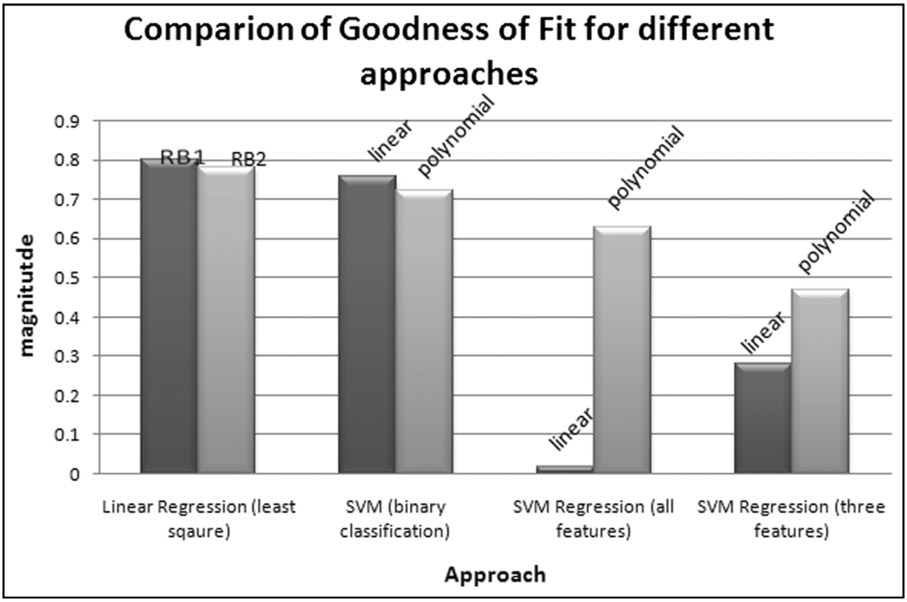
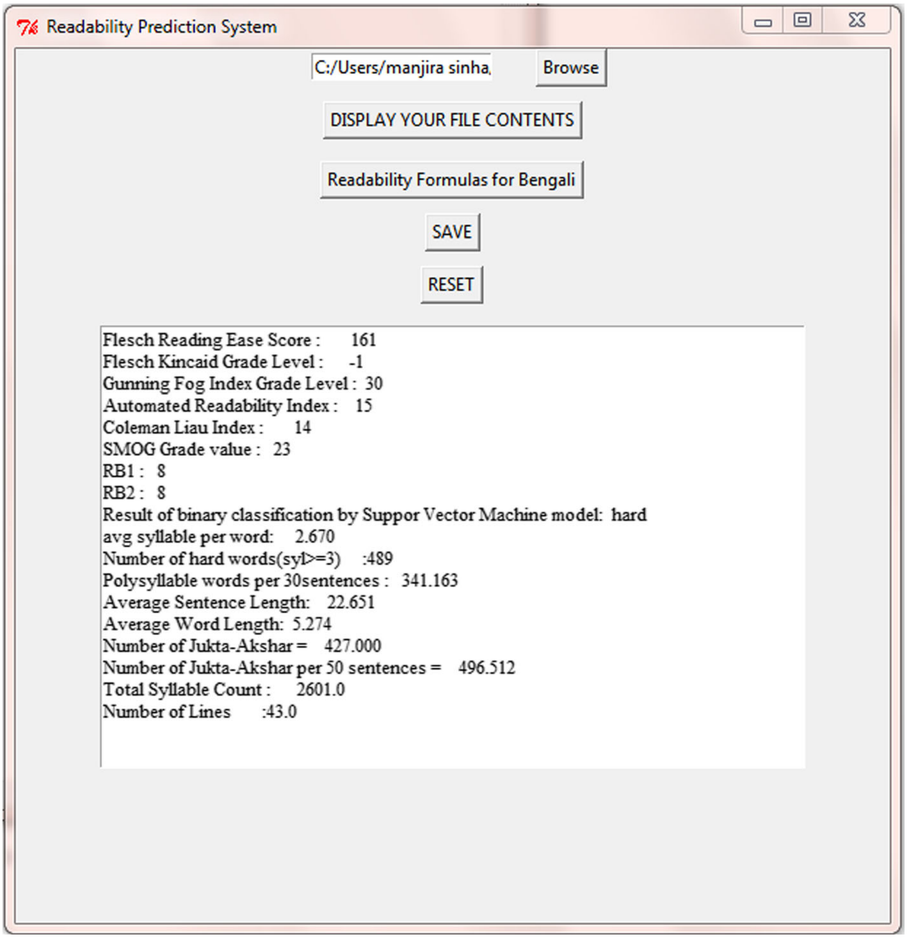


Fig. 10 Comparison of goodness of fit for the different proposed readability models

comparable R^2 values. Whether to use regression or classification in determining the *hardness* of a text depends on the intended application. Moreover, from Figs. 9 and 10, it can also be inferred that, although SVR models with polynomial kernel have comparable R^2, they incur large errors than linear regressions.

1.10.1 Readability prediction system

In the introduction, we have declared that our aim is to incorporate the models and findings from our study into an input–output system that can be easily used for measuring the difficulty of a given Bangla document. To achieve this goal, we have developed a system (refer to Fig. 11 below), which given a Bangla text document provides as outputs the values of the text attributes as well as predictions by two Bangla regression models and the result from binary SVM classification. SVR has not been incorporated due to its poor performance as described above. The scores from the English readability formulae have also been incorporated for the sake of transparency.



Readability Prediction System
Browse
DISPLAY YOUR FILE CONTENTS
Readability Formulas for Bengali
SAVE
RESET

C:/Users/manjira sinha,

```
Flesch Reading Ease Score :     161
Flesch Kincaid Grade Level :    -1
Gunning Fog Index Grade Level :  30
Automated Readability Index :   15
Coleman Liau Index :      14
SMOG Grade value :  23
RB1 :  8
RB2 :  8
Result of binary classification by Suppor Vector Machine model:  hard
avg syllable per word:    2.670
Number of hard words(syl>=3)     :489
Polysyllable words per 30sentences :   341.163
Average Sentence Length:   22.651
Average Word Length:  5.274
Number of Jukta-Akshar =    427.000
Number of Jukta-Akshar per 50 sentences =    496.512
Total Syllable Count :     2601.0
Number of Lines       :43.0
```

Fig. 11 Snapshot of the first version of the readability prediction system

At present, our system does not contain any provision to incorporate the specific user backgrounds. Currently we are expanding the study to different user groups and subsequently the system will able to predict the relative difficulty of a text depending on the user characteristics.

1.11 Conclusion and perspective

In this paper, we have presented different techniques to predict comprehension difficulty of Bangla texts. We have used syntactic and lexical text features, majority of which have not been used earlier to model the text readability in Bangla. Through regression, we have established two readability models for Bangla based on average word length in terms of visual units, number of polysyllabic words and number of jukta-akshars. Similarly, for SVM, we have shown that linear and polynomial kernel functions deliver the best results in predicting the text difficulty; SVM predicts the class label (easy or hard) of a text based on all of the features. We have also combined the two approaches in support vector regression (SVR) method, but SVR has been found to perform less accurately than traditional regression technique. In the course of the paper, we have shown that although the classical English readability formulas correlate with the user evaluation, they are not helpful as text difficulty predictors. The study of Bangla readability (Das and Roychoudhury 2006) has studied comparison of one and two parametric fits through regression for a miniature model, but they have not considered parameters like AWL, JUK; we have found these parameters among the major players. Moreover, according to our regression analysis, any one-parametric hyperbolic fit other than the SMOG equivalence is not appropriate as they generate negative values of R^2. According to the best of our knowledge, this is the first work in Bangla, which attempts to develop text readability measures using regression and machine learning methods, and compares among them. In future, we plan to extend the binary SVM to a multi-class classifier (e.g., easy, moderate, hard) and augment the feature space with more text properties such as part of speech statistics. We are also working on developing devising readability predictors targeted towards different age group such as school students and different social groups such as the first generation learners from economically backward strata. We will also like to study the effects of other textual features like semantics, coherence etc. on readability of Bangla Texts.

References

Agnihotri, R. K. (2008). 13 orality and literacy. *Language in South Asia*, page 271.

Bamberger, R., & Rabin, A. T. (1984). New approaches to readability: Austrian research. *The Reading Teacher, 37*(6), 512–519.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews, 11*(10), 203–224.

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

Britton, B., & Gülgöz, S. (1991). Using kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*(3), 329.

Buswell, G. (1937). *How adults read*. University of Chicago.

Chakraborti, P. (2003). *Diglossia in Bengali*. PhD thesis, University of New Mexico.

Chall, J. (1958). *Readability: An appraisal of research and application*. Number 34. Ohio State University.

Chall, J. (1995). *Readability revisited: The new Dale-Chall readability formula, volume 118*. Cambridge: Brookline Books.

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

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Cotugna, N., Vickery, C., & Carpenter-Haefele, K. (2005). Evaluation of literacy level of patient education pages in health-related journals. *Journal of Community Health, 30*(3), 213–219.

Crossley, S., Dufty, D., McCarthy, P., & McNamara, D. (2007). Toward a new readability: A mixed model approach. In *Proceedings of the 29th annual conference of the Cognitive Science Society*, pp. 197–202.

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

Das, S., & 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.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems, 9*, 155–161.

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

DuBay, W. (2007). *Smart Language: Readers, Readability, and the Grading of Text.* ERIC.

Ferguson, C. A. (1959). Diglossia. *Word-Journal of the International Linguistic Association, 15*(2), 325–340.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221.

Foltz, P., Kintsch, W., & Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*(2–3), 285–307.

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

Graesser, A., McNamara, D., & Kulikowich, J. (2011). Coh-metrix providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234.

Graesser, A., McNamara, D., Louwerse, M., & 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*. NewYork: McGraw-Hill.

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*, (pp. 71–79). Association for Computational Linguistics.

Islam, Z., Mehler, A., Rahman, R., and Texttechnology, A. (2012). Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*.

Kemper, S. (1983). Measuring the inference load of a text. *Journal of Educational Psychology, 75*(3), 391.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Technical report, DTIC Document.*

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

Klare, G. (1963). *The mesaurement of readability.* Ames: Iowa State University Press.

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

Learning, R. (2001). *The atos readability formula for books and how it compares to other formulas*. Madison: School Renaissance Institute.

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*, (pp. 548–549). ACM.

Manning, C. D., Raghavan, P., & 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.

McNamara, D., Louwerse, M., McCarthy, P., & Graesser, A. (2010). Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292–330.

Miltsakaki, E., & Troutt, A. (2007). *Read-x: Automatic evaluation of reading difficulty of web text.* In *Proceedings of E-Automatic evaluation of reading difficulty of web text.* In *Proceedings of ELearn*.

Montgomery, D., Peck, E., and Vining, G. (2007). *Introduction to linear regression analysis*, volume 49. Wiley.

Oakland, T., & Lane, H. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing, 4*(3), 239–252.

Petersen, S. E., & 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*, (pp. 46–76).

Rosch, E. (1978). *Principles of categorization. Fuzzy grammar: a reader* (pp. 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*, (pp. 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., & Callan, J. (2003). A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems (TOIS), 21*(4), 457–491.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199–222.

Stenner, A. (1996). *Measuring reading comprehension with the lexile framework.*

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology Section A, 57*(4), 745–765.

vor der Brück, T., Helbig, H., Leveling, J., & Kommunikationssysteme, I. (2008). *The Readability Checker Delite: Technical Report*. FernUniv., Fak. für Mathematik und Informatik.

Zar, J. (1998). Spearman rank correlation. *Encyclopedia of Biostatistics.*