A comparative study of feature selection methodologies in a readability assessment framework for children’s literature

THESIS

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Ritu Singh

Graduate Program in Computer Science and Engineering

The Ohio State University

2012

Master's Examination Committee:

Prof. Eric Fosler-Lussier, Advisor

Prof. Rajiv Ramnath
Abstract

Assessment of the reading level in children’s literature is often desired by the educational community, parents and writers in designing a better curriculum to facilitate a child’s educational growth. A comprehensive readability assessment framework that can provide precise and fine-grained leveling of children’s books is desirable. We have conducted a study within the readability assessment framework developed by Ma et al. (2012a): in our work, we perform a comparative analysis of feature selection methodologies in order to better understand the relative importance of features and their relevance in predicting readability in terms of a standard readability assessment framework. While performing data analysis, we find that different features show a number of patterns of variation across levels. A better understanding of the nature and importance of these features based on level can help in identifying some orthogonal directions that can help eventually to build a vector-based metric of readability. We find that there are some features that have a negative impact on the system’s ability to predict reading level. However, with our feature selection methodologies, we are not able to get statistically significant improvement in prediction performance of our standard readability assessment framework (Ma et al. 2012a).
Acknowledgments

I would like to thank my parents and brother for their support and guidance in every phase of my life.

I would like to express my deep and sincere gratitude to my advisor Prof. Eric Fosler-Lussier and to Prof. Rajiv Ramnath for their guidance all through the work.

I would like to thank Robert Lofthus from Xerox Corporation and my colleague Yi Ma and other SLATE lab members at OSU for their cooperation and input. This work was supported by a collaborative project sponsored by Xerox.
Vita

Oct 1987 ...................................................... Born in Jind, India

Jun 2010 ...................................................... B. E. in Information Technology, University of Delhi

Sep 2011- Mar 2012 ...................................... Graduate Fellow, The Ohio State University

Apr 2012- Jul 2012 ................................. Graduate Research Associate, Speech and Language Technologies Lab, The Ohio State University

Publications


Fields of Study

Major Field: Computer Science and Engineering
# Table of Contents

Abstract ...................................................................................................................... ii

Acknowledgments .................................................................................................. iii

Vita .......................................................................................................................... iv

Table of Contents .................................................................................................... v

List of Tables .......................................................................................................... vii

List of Figures ......................................................................................................... viii

1. Introduction ........................................................................................................ 1
    1.1. Motivation and Background ..................................................................... 1
    1.2. Problem Statement .................................................................................. 6
    1.3. Goals of the work .................................................................................... 7
    1.4. Scope and limitation of the work ............................................................. 8
    1.5. Key Terminology ..................................................................................... 9

2. Literature Survey ................................................................................................ 10
    2.1 Different methodologies of readability assessment .................................... 10
    2.2. A review of different methodologies of feature selection ....................... 15

3. Methodology ...................................................................................................... 18
    3.1. Approach ................................................................................................. 18
3.1.1. Data Preparation: ................................................................. 18

3.1.2. Features: ............................................................................. 19

3.2. Data Analysis: ......................................................................... 21

3.3. First experimental study ............................................................ 36

3.3.1. Feature selection- Forward feature optimization ..................... 36

3.3.2. SVM-rank machine learning algorithm .................................... 38

3.4. Second experimental study: Multi-level classification of features: ........ 40

4. Substantiation of the readability assessment framework using different approaches 61

4.1. Performance results using SVM-ranking ...................................... 61

4.2. Results using multi-level classification ........................................ 70

5. Conclusion and Remarks ................................................................ 73

References ...................................................................................... 74
List of Tables

Table 1: Prediction accuracy with single feature ................................................. 62
Table 2: Prediction accuracy for the best five features i.e., feature 1,3,7,11,12 and with
the addition of one single feature at a time from the rest of the features ..................... 63
Table 3: Prediction accuracy of the six best features with an addition of one feature from
the remaining subset of features ................................................................................ 63
Table 4: Prediction accuracy of the seven best features with the addition of one feature
from the remaining subset of features ......................................................................... 64
Table 5: Prediction accuracy of the eight best features with the addition of one feature
from the remaining subset of features ........................................................................ 64
Table 6: Prediction accuracy of the nine best features with the addition of one feature
from the remaining subset of features ......................................................................... 64
Table 7: Prediction accuracy of the ten best features with the addition of one feature from
the remaining subset of features ................................................................................ 65
Table 8: Prediction accuracy of the eleven best features with the addition of one feature
from the remaining subset of features ......................................................................... 65
Table 9: Prediction accuracy of all 14 features with training dataset ............................. 66
Table 10: Results on evaluation data set with completely new data set of 38 books ........ 66
Table 11: Results using multi-level classification ......................................................... 70
List of Figures

Figure 1: Number of words in entire book versus level of the book ...................... 22
Figure 2: Number of letters per word versus level of the book .......................... 23
Figure 3: Number of sentences in entire book versus level of the book ................. 24
Figure 4: Number of words per sentence versus level of the book ........................ 25
Figure 5: Type token ratio versus level of the book ........................................... 26
Figure 6: Number of pages versus level of the book ......................................... 27
Figure 7: Number of words per pdf page versus level of the book ...................... 28
Figure 8: Number of sentences per page versus level of the book ....................... 29
Figure 9: Number of textlines per page versus level of the book ......................... 30
Figure 10: Number of words per textline versus level of the book ....................... 31
Figure 11: Number of words per textbox versus level of the book ....................... 32
Figure 12: Number of textlines per textbox versus level of the book .................. 33
Figure 13: Average ratio of textbox area to page area versus level of the book ........ 34
Figure 14: Average font size versus level of the book ....................................... 35
Figure 15: Number of words in entire book versus level of the book .................... 42
Figure 16: Number of letters per word versus level of the book .......................... 44
Figure 17: Number of sentences in entire book versus level of the book ................ 45
Figure 18: Number of words per sentence versus level of the book ...................... 46
Figure 19: Type token ratio versus level of the book ........................................... 47
Figure 20: Number of pages versus level of the book ....................................... 48
Figure 21: Number of words per page versus level of the book ........................................ 49
Figure 22: Number of sentences per page versus level of the book ................................. 51
Figure 23: Number of textlines per page versus level of the book .................................. 52
Figure 24: Number of words per textline versus level of the book ................................. 53
Figure 25: Number of words per textbox versus level of the book ................................. 54
Figure 26: Number of textlines per textbox versus level of the book .............................. 55
Figure 27: Average ratio of textbox area to page area versus level of the book ............... 56
Figure 28: Average font size versus level of the book ................................................. 57
1. Introduction

1.1. Motivation and Background

Readability characterizes the ease of reading, understanding and comprehending a text. Assessment of reading level of children’s literature is one of the common challenges faced by the educational community. Knowledge of the reading level of children’s books is used in a variety of contexts: it helps parents in selecting appropriate books for their children; it provides help to teachers in the leveling of books in the school library; and it gives guidance to writers writing for different literacy needs (for example, text simplification during authoring process). Readability assessment research has been an on-going stream of research on reading and text complexity in other areas such as Second Language Acquisition and psycholinguistics.

The selection of books and study material is not only based on the grade level of the child but is more geared towards the ability of the child, and this carries a lot of responsibility from a teacher’s perspective to have a proper knowledge of levels of books before guiding a child. The educational community typically makes fine-grained distinctions in reading levels, with each grade being covered by multiple levels. These finer distinctions of grade have a significant contribution in guided
reading: guided reading has proven to provide significantly beneficial effects on helping students develop reading skills (Avalos et. al., 2007). Guided reading is one of the most effective tools to improve a student's fundamental reading skills and to help the student develop higher level comprehension skills. Different levels per grade cater to the different educational abilities of children. There are different leveling mechanisms that may take into account multiple factors; for example, 1st grade is covered, approximately, by levels 3–14 on the Reading Recovery scale (M.M. Clay, 1993), or levels C to H in the Fountas and Pinnell leveling system (Fountas and Pinnell, 2010). For our readability assessment framework, we are considering Fountas and Pinnell reading levels as the gold standard.

Traditional methods of readability assessment provide a gross index of difficulty; these readability formulas were based on countable features of text, like number of syllables (Flesch, 1948). These formulae are suitable only for providing a rough estimate of grade level. Most of the earlier formulae failed to include factors related to the interaction between a reader and the text. Various studies have shown that these formulae do not address some significant factors, such as content, visual input, purpose, and organization. Even at higher levels, teachers find it hard to rank the reading ease of texts. For this reason, a lot of research studies have been done for the better assessment of readability.

With recent advancements in areas of cognitive sciences and artificial intelligence, a better readability assessment framework is desired which can provide higher levels of prediction accuracy with respect to the finer-grained levels within a grade and can
provide a handy tool for teachers for guided reading. For determining coarser grade levels, common machine learning techniques like classification and regression have proved to be quite promising (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005; Peterson and Ostendorf, 2005; Feng et. al., 2010). In the recent study conducted by Ma et al. (2012a), it has been shown that the traditional machine learning techniques, like classification and regression, have given the worst performance in predicting fine-grained tasks. Earlier approaches have not dealt with problems when there are an increased number of classes and when the reading levels are not linearly distributed. An increased number of classes does not allow classification techniques to succeed with a more fine-grained leveling system. Even regression techniques have not given good results as these finer-grained reading levels are not linearly distributed. In the framework developed by Ma et al. (2012a), a ranking-based approach gave better results on a fine-grained book-leveling problem for kindergarten through third-grade books than either classification or regression. They found ranking methodology to be a better approach when there is a non-linear reading level scale.

With advancements in the areas of natural language processing (NLP), a significant number of features has been considered to provide a better readability assessment (Si and Callan, 2001; Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, Chapter 1. Introduction 2 2005; Heilman et. al., 2007, 2008; Pitler and Nenkova, 2008; Barzilay and Lapata, 2008; Petersen and Ostendorf, 2009). The above approaches utilized mostly statistical analysis of a text at lexical and syntactical level.
and were quite successful in giving good performance for readability assessment for higher grade levels. However, few studies have been done in early primary children’s literature: these advanced features, such as syntactic patterns, did not contribute much to the performance in predicting reading level for these finer-grade levels of children’s literature (Ma et al., 2012a). It is likely that high-level text-based features carry more importance for higher grade levels as compared to children’s literature. In our earlier studies, we have shown that visually-oriented features (e.g., ratio of text region area to whole page area, font size, etc.) carries higher importance in predicting the reading levels of children’s books where pictures and textual layout dominate the book content over text. In our current study, besides text-based features, we have also considered the features based on visual layout of the page. Capturing information pertaining to visual layout is not an easy process. Researchers who account for these features for the readability assessment generally use human participants. In our readability assessment framework, we were quite successful in automating this process of predicting reading level with much less human intervention (Ma et al., 2012b).

However, when we extended our data set to improve the performance of our readability assessment framework, we have found that different features show a number of patterns of variation across levels. A better understanding of the nature and importance of these features based on level can help in creating a better conceptual layout for the system. We have tried different methodologies for feature selection to
come up with a better readability assessment framework as well as understand the contributions of features to overall performance.

Feature selection allows us not only to better understand the domain, but it has some added advantages when we have a larger data set. Having fewer, more reliable and relevant features can give better results. While doing feature selection, we can hope to get a safe "number of dimensions" which are needed for optimal classification for a given size of data-set. Moreover, with a reduced set of features, we can reduce the complexity of the system. A specific subset of features may give more accuracy.
1.2. Problem Statement

In our study, we are trying to enhance the capability of a readability assessment framework (Ma et al., 2012a) by doing a sensitivity analysis of the feature space in order to understand the relative importance of features which can help in laying a better conceptual layout of the framework. A detailed analysis of feature space is required to find a set of relevant features and to identify some orthogonal directions that can help eventually to build a vector-based metric of readability.

In our study, we are doing a comparative analysis of different feature selection methodologies to better understand the nature and contribution of different features across levels of books for predicting their reading level in a readability assessment framework.
1.3. Goals of the work

The following are the major goals of this comparative analysis of feature selection methodologies in the context of our readability assessment framework:

- Understand the nature of the features which are most relevant in enhancing the performance of the framework.
- Identify some orthogonal directions that can eventually build a vector-based metric of readability.
- Understand the possible directions of classifying features when we have a non-linear reading level scale.
- Better understand the domain of readability assessment of children’s literature.
- Find a better way of analyzing feature space and providing new directions for analyzing features important for readability assessment of children’s books.
1.4. Scope and limitation of the work

The scope of this study is limited to the readability assessment framework developed by Ma et al. (2012a) for children’s literature. We are considering only the features which are most pertinent to elementary education. Different new features become more important when we analyze the higher grade levels. But our study is limited to our framework for children’s literature, although we hope that our comparative analysis of methodologies of feature selection can be used in a variety of contexts.
1.5. Key Terminology

Here we define some of the key terms we followed while writing this document and portraying our understanding of the terms during this study.

- **Feature:** A feature is the specification of an attribute and its value. We have used feature as a synonym for attribute (e.g., in feature-subset selection). And in this document, we have used variable and feature interchangeably.

- **Feature vector:** A list of features describing an instance (in our context, an instance is a book).

- **Class:** We have defined class as a combination of levels of books which can be separated on the basis of mean values of a feature. For example, one class $C_1 = L[1,2,3]$ implies class $C_1$ is a combination of levels 1, 2 and 3.

- **Level:** A level is defined as the level of a book according to the Fountas and Pinnell benchmark assessment system 1 (Fountas and Pinnell, 2010). A grade consists of some levels. For example, grade 2 is equivalent to guided reading levels J through M.
2. Literature Survey

2.1 Different methodologies of readability assessment

Readability refers to the difficulty or ease of reading a given text. There are quite similar but still distinguishable definitions about readability. George Klare (1963) defines readability as “the ease of understanding or comprehension due to the style of writing.” His definition has put more stress on the writing style as compared to the text content. Gretchen Hargis and her colleagues at IBM (1998) emphasized the clarity of a text by stating that readability is the “ease of reading words and sentences.” Edgar Dale and Jeanne Chall (1949) defined readability in a more comprehensive manner as “The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success 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.”

The history of readability assessment dates back to the 1920s when Thorndike (1921) measured text difficulty by tabulating words according to the frequency of their usage in general literature. Since then, a lot of research has been done with the main focus on arithmetic metrics. Most of the readability formulas were based on those features of text that can be subjected to mathematical calculations. The Gunning Fog Index
(The Technique of Clear Writing, Robert Gunning, 1952), the Dale-Chall formula (1948), the Flesch Scale (A New Readability Yardstick, Rudolph Flesch, 1948), the Fry Graph (1977) and the Spache (1974) were some of the famous readability formulas which have used arithmetic metrics mostly based on word difficulty and sentence difficulty measures. These readability formulas were widely used in research, healthcare, law, insurance, and journalism. The U.S. military has developed its own set of formulas for technical training materials (William H. DuBay, 2004). By the 1980s, there were around 200 formulas, and over a thousand studies on the readability formulas have been conducted. These formulas served as more of a quantitative and objective tool for assessing the readability of a text (C. Analytics, 2004).

Earlier research laid more importance on the assumption that text readability assessment is a simple function of some shallow text-based properties (Miller and Kintsch, 1980). These formulas have given more stress on semantic (e.g., the vocabulary) and syntactic factors (e.g., number of characters in a sentence) of a text. These formulas focused more on the text-based features; they failed to include factors which reflect the interactive nature of the text: for example, the visual layout of the text, which turned out to be one of the important factors for assessment of reading level of children’s literature (Ma et al., 2012a).

Major factors considered in determining the difficulty (or ease) while assessing readability are the vocabulary level; the structure of the sentence; print factors, like graphical aids, pictures, font size; the level of comprehension required; etc. The
importance of considering these factors while assessing reading level varies according to the grade level of text. For example, the graphic aids are more important for children’s text in comparison to higher grade levels.

Advancements in computational linguistics have made it possible to automatically extract a wider range of language features from text. Varied advanced machine learning techniques have been utilized, which involves development of language models and parsers that can explore more complex lexical features and syntactic constructs in aiding readability study. Different independent features are explored by recent statistical methods. Si and Callan (2001) have used language model to capture the content information. Collins-Thompson and Callan (2004) have demonstrated that reading difficulty can be estimated with a simple language modeling approach using a modified naïve Bayes classifier. They have shown that the Smoothed Unigram method is quite robust in the analysis of web documents and passages. Schwarm and Ostendorf (2005) and Petersen and Ostendorf (2009) both used a support vector machine to classify texts based on the reading level. They combined traditional methods of readability assessment and the features from language models and parsers. Heilman et. al. (2007) studied the impact of grammar-based features combined with a language modeling approach for readability assessment of first- and second-language texts. They argued that grammar-based features are more pertinent for second-language learners than for the first-language readers. Aluisio et. al. (2010) have developed a tool for text simplification for the authoring process and to predict the readability level corresponding to the literacy level of the target reader. They
addressed lexical and syntactic phenomena to make text readable, but their assessment takes place at more coarse levels of literacy, i.e., rudimentary, basic or advanced; instead of finer-grained levels used for children’s books.

A detailed analysis of various features for automatic readability assessment has been done by Feng et. al. (2010). They studied various kinds of features, including shallow, language modeling, parts of speech, syntactic, and discourse features; and evaluated their impact on predicting grade level of reading material for primary school students. Most of the previous work has used web page documents, short passages or articles from educational newspapers as their datasets; typically the task is to assess reading level at a whole-grade level. In contrast, early primary children’s literature is typically leveled in a more fine-grained manner. With the advancements in education, the educational community started classifying each grade into more fine-grained levels. With the introduction of guided reading and reading recovery projects (R.J. Tierney, J.E. Readence, 2000), educators started desiring a better readability assessment framework that can level books at a much fine-grained level. A guided reading approach was found to be very helpful in increasing the reading level of children. The guided reading approach recognized that a wide range of reading abilities exists within any grade level, and that reading at the appropriate levels ensures success. Recently, different kinds of methodologies have been used to provide the level of children’s books. Following are some of the common leveling systems used for children’s literature:

- Grade Level: the most basic system with sorting by grade level.
• Guided Reading level: developed by Irene Fountas and Gay Su Pinnell, and they called their system *Leveled Literacy Intervention (LLI)*. They provided a benchmark assessment system for grade levels K-8th. They provided a guided reading level system – several levels within each grade level – which gives a more precise reading level for books. For example, grade 2 is equivalent to guided reading levels J through M. In our readability assessment, we are using their benchmark assessment system as the gold standard (Fountas and Pinnell, 2010).

• The Lexile Framework for reading: assesses a book’s difficulty and helps match reader ability and text difficulty based on the numeric Lexile scale. The underlying algorithm is based on measuring vocabulary and sentence length, and the system targets books at the right reading level for the child’s ability (C. Lennon and H. Burdick, 2004).

• Developmental Reading Assessment: a reading assessment tool intended to identify the independent reading level for students in grades K–8. This tool measures reading accuracy, fluency and comprehension (J. Beaver and M.A. Carter, 2001).

• Reading Recovery: an intensive one-on-one remediation program designed to supplement reading instruction for students in grades K–2 who are slow to read (M.M. Clay, 1993).
2.2. A review of different methodologies of feature selection

Feature selection has been an area of interest in a variety of contexts. It is the methodology of selecting a subset of relevant features for building robust learning models. Some people distinguish between variable and feature selection. They consider “variable” as the raw input variables and “features” variables constructed for the input variables (Guyon and Elisseeff, 2003). Feature (or variable) selection is one of the common techniques for learning data in fields like biology, where there are thousands of features to be considered, but only a few studies have been done in areas of text classification. Guyon and Elisseeff (2003) have given an overview of the different methodologies of variable and feature selection used in different areas. Forman (2008) has demonstrated different techniques of feature selection in the areas of text classification. Chapelle and Sathiya Keerthi (2008) have studied feature selection in a multi-class setting and multi-labeled text classification, where they developed an embedded method to find a small set of features for all the classes simultaneously. Although a lot of work has been done in areas of feature selection, very limited work has been done in areas related to classifying features in children’s literature. In a broader sense, feature selection methods can be classified into the following types:

- Filtering: typically involves simple techniques for weeding out irrelevant features without fitting a model. It is usually not an efficient approach, but it helps in removing features which are too irrelevant. Usually this approach is
applied in the beginning of feature selection. It doesn’t consider the type of algorithm to be considered for learning data.

- **Embedded methods:** involve consideration of kernel methods where feature vector space is mapped into a higher dimensional feature space. Expansion of feature space provides potentially useful features. Though some people classify it under feature selection algorithm, it does not really do a feature selection.

- **Wrapper methods:** include methods where they find features that work best with some particular learning algorithm:
  
  - Brute-force approach: Trying all the possible combinations of features. It usually involves the selection of features which perform best on the test set. It is a very expensive and generally unfeasible approach and has a danger of overfitting.
  
  - Forward stepwise selection: It involves the addition of features one by one and works best performance-wise in the learning algorithm.
  
  - Backward elimination method: It involves the stepwise removal of the feature which is performing worst in the learning algorithm.

Wrapper methods require running a base learning algorithm many times, and that can be very time consuming and expensive to run.
Most learning algorithms implicitly do feature selection. For example, different algorithms, like support vector machines, use maximum margin hyperplane to focus on the important features and ignore the irrelevant features. But still we need feature selection to better understand the underlying model and probably for improved accuracy. In our readability assessment framework, we might not be sure of getting improved performance, but with this study we are hoping to get a better understanding of features and their role at a particular level of a book. We are trying to understand which features are relevant to the system and how they are correlated.

There are some limitations to feature selection algorithms.

- If a feature is not selected during the feature selection algorithm, it does not mean that feature is not a strong predictor.
- It may throw away some features which are desired by the domain expert.
- Most feature selection methods are greedy and might not find an optimal feature set.
- Feature selection can overfit the data.

But as a whole, doing feature selection analysis provides a better understanding of the system.
3. Methodology

3.1. Approach

3.1.1. Data Preparation:

While designing a readability assessment framework for children’s literature, we have used 97 books in our training data set and 38 books in the evaluation data set (an expansion of the data set found in Ma et al (2012b)). The reading levels of these books range from level A to N (i.e., 1 to 14) in the Fountas and Pinnell Benchmark Assessment system 1 (Fountas and Pinnell, 2010), which served as our gold standard. These books covered a range of diverse genres, series and publishers. The level of these books corresponds roughly with kindergarten to third grade, thus covering mostly the primary grade books with increasing level of difficulty from A to N.

We started with the physical copies of books rather than electronic versions due to agreement constraints with the book’s publishers. We scanned each book into a PDF version followed by running OCR (Optical Character Recognition) using Adobe Acrobat. This was followed by automatic annotation of each book using Adobe Acrobat and extracting corresponding XML for each book. The OCR process introduced a lot of errors. In our
prior experiments (Ma et al., 2012b), we have compared human annotation, which removes these errors manually to obtain the correct text for feature extraction, with automatic annotation, which keeps these OCR errors but reduces a lot of human intervention. In the results of that study, we did not observe a significant performance drop in predicting the reading level for scanned children’s books. For our current study, we have utilized the automatic annotation process followed by feature extraction from the XML files obtained after annotating the PDF files. The automatic annotation process and XML generation process is implemented as a JavaScript plug-in menu item within Adobe Acrobat. The feature set which we have considered for our experiments will be explained in the next section.

3.1.2. Features:

In our readability assessment framework, we have considered two types of features for our feature selection methodologies:

- Surface-level features:
  - Number of words in the entire book
  - Number of letters per word
  - Number of sentences in the entire book
  - Number of words per sentence (average sentence length)
  - Type-token ratio of the text content: Type-token ratio measures the vocabulary variation within a text.

- Visually-oriented features:
- Number of PDF pages (page count)
- Number of words per page
- Number of sentences per page
- Number of textlines per page
- Number of words per textline
- Number of words per annotated text rectangle
- Number of textlines per annotated text rectangle
- Average ratio of annotated text rectangle area to page area
- Average font size

We have called the latter features visually-oriented features because we need the visual layout of the book for obtaining these features.
3.2. Data Analysis:

We performed a sensitivity analysis with different approaches to examine variations in features with different levels of books. The main purpose for feature selection is to understand what features are contributing to performance and which are irrelevant. We wanted to determine the relative importance of features and, hence, to find a vector measure of readability. We have plotted the following graphs to understand how feature values vary with the levels of books in our training data set:

On the x-axis (horizontal axis), we are plotting the feature values, and

On the y-axis (vertical axis), we have the levels of books (1-14) – these levels correspond to Fountas and Pinnell levels A-N.

With the analysis of these graphs, we wanted to see what possible techniques we can use to have a better understanding of the feature space.
Figure 1: Number of words in entire book versus level of the book

Feature 1: Number of words in the entire book: This feature gives the total number of words in the book. In a general sense, the number of words in the entire book increases with the level of the book. But as seen in this graph, there are some outliers. For example, one book having level 11 (Level K according to F&P) has a quite high number of words, even more than books having level 12, 13, or 14.
Feature 2: Number of letters per word gives the average number of letters in a word per book. It has been found that usually lower grade level books have few letters per word. But there are quite a few exceptions to this. Hence, we can see a lot of outliers in this graph. For example, the words like ‘Christmas’ or ‘Television’ are very common words that can be found in lower level books. Moreover, while preparing the dataset, we saw that in some books, the main subject’s name itself consists of many letters, like ‘Margaret’. These frequent words in such books create a lot of outliers.
Figure 3: Number of sentences in entire book versus level of the book

Feature 3: Number of sentences in entire book

As can be seen in the graph, the average number of sentences generally increases with the level of the book.
Figure 4: Number of words per sentence versus level of the book

Feature 4: Number of words per sentence.

It is the average number of words per sentence.
Feature 5: Type token ratio: It measures the vocabulary variation within a text. Its data values are quite widespread and do not show a proper trend with an increase in the level of the book.
Feature 6: Number of pdf pages in the book.

In general, the number of pages increases with an increase in the level of the book.
Feature 7: Number of words per pdf page.

It is the average number of words per page, and it generally increases with the level of the book.
Figure 8: Number of sentences per page versus level of the book

Feature 8: Number of sentences per pdf page.

It is the average number of sentences per page.
Feature 9: Number of textlines per pdf page.

It is the average number of textlines per page. As can be seen from the figure, this feature generally increases with the level of the book.
Figure 10: Number of words per textline versus level of the book

Feature 10: Number of words per textline.

It is the average number of words per textline.
Feature 11: Number of words per textbox.

It is the average number of words per textbox. A textbox is a feature generated after running automatic annotation with the JavaScript plug-in in Adobe Acrobat. It generates a rectangular textbox around a continuous nearby text (a distance threshold of 22 pixels is used, Ma et.al., 2012b).
Feature 12: Number of textlines per textbox.

It is the average number of textlines per textbox generated by automatic annotation.
Figure 13: Average ratio of textbox area to page area versus level of the book

Feature 13: Average ratio of textbox area to pdf page area. It is the average of ratio of textbox area to page area.
Feature 14: Average font size.

It is the average font size used in the book. As can be seen from the graph, generally font size is large for lower level books.
3.3. First experimental study

3.3.1. Feature selection- Forward feature optimization

Feature selection has many potential benefits. Our focus in this study is to find a subset of features that are contributing to the performance and to eliminate the irrelevant features. Selecting the features requires proper domain knowledge. But how should we proceed when we suspect some interdependence of features? A feature performing badly in isolation can contribute to (or enhance) the performance of the system when combined with other features. There can be some irrelevant and redundant features. Different features have different relevance when we consider statistical significance. With feature selection, we hope to improve the performance (i.e., increase predictive accuracy) and, hence, can enhance the understanding of the underlying concepts in the model. We use a predictor to measure the relevance (i.e., accuracy). Variable selection can be done using different approaches (for details refer to section 2.2):

(a) exhaustive evaluation of the possible feature sub-sets,

(b) elimination of the worst feature (backward elimination),

(c) forward selection of feature (selecting the best feature at each stage)

For our experiments, we adopted the third method. In the forward feature optimization method, we did feature selection according to the following steps:
• Start with a single feature and do the performance analysis.
• Select the feature giving the best performance and repeat the experiments with the rest of the features.
• Again, select the feature giving the best performance and add it to the feature subset. We repeat the experiments till we analyze all the features or no further performance improvement is seen.

This method is a greedy method and it has its own advantages and limitations.

Advantages of the proposed feature selection method:

(a) It gives improved accuracy on the training set. Since we are selecting the feature which gives the best performance locally (greedy approach), we are optimizing our current results.

(b) It gives an easy to understand and less complex model.

(c) We can discard the rest of the features, which reduces the predictive performance of the model.

(d) It maximizes the relevance of a subset of features.

Limitations of the forward-feature optimization feature selection method:

(a) It is a greedy approach, which might not always give an optimal feature set and, moreover, it ignores the relationship between features.
(b) This feature selection method has a tendency to overfit the data when not enough data is available.

(c) A feature which is useless by itself can give better performance with other features.

3.3.2. SVM-rank machine learning algorithm

We have done our experiments using the readability assessment framework proposed by Yi Ma (Ma et al., 2012). In this algorithm, we classify books by first ranking all of the books and then assigning the level based on the ranking output. We used the SVM$^\text{rank}$ ranker (Joachims, 2006) for our experiments. We prepared the data set using the methodology described in Section 3.1. We used the features obtained from XML files of each book as input training data in the SVM ranker. Our SVM-rank machine learning algorithm works as:

- The SVM$^\text{rank}$ ranker first learns from the training dataset and then generates a ranking output for each training dataset.
- Based on the ranking output, we sort our training dataset.
- During the testing phase, the SVM ranker generates a ranking output when we input a feature vector for a test book. Based on the ranking output, the ranker places the test book according to the rank predicted by the trained ranking model.
• Then the reading level of the test book can be inferred from the reading levels of neighboring leveled books in the rank-ordered list of books. The level of the book is predicted by averaging the levels of three books on its left and three books on its right weighted by distance, i.e., the book which is adjacent has more weightage in determining the level of the test book.

Experimental results using this methodology are given in Section 4.1.
3.4. Second experimental study: Multi-level classification of features:

In our second approach, we have used a simple multi-level classification method for learning data. Support vector machines are good at two-class discriminative learning problems. But while analyzing the data, we found that for some features, the average data values are nonlinear with respect to the levels and therefore can be divided into multiple classes for each feature. The data is such that it can’t be classified into different levels of books as classes, i.e., the data is not mutually-exclusive per level of the book. Hence we analyzed another method which clubs the non-mutually exclusive data into classes where a class has many levels of books. Multi-level classification can be seen as a modified version of multi-label classification (G. Tsoumakas and I. Katakis, 2007) with a combination of multi-class classification (M. Aly, 2005). To each instance, we assigned different levels depending on the class.

In our readability assessment algorithm, Multi-level classification is done in the following steps:

- We divide the data set of each feature into multiple possible classes of levels of books.

- Classification of feature values into different classes is based on the difference in mean values of feature values. Mean values are calculated per level for each book. Based on the proximity of mean values per level, different classes are defined. For example, one feature can have seven classes based on the distribution of data, whereas another feature might have four distinguished classes.
• The difference in the mean values defines the boundary values for each class: for example, if m1 is the mean of level 1 and m2 is the mean of level 2 and the difference between these two levels is more than a threshold (where the threshold is calculated based on the minimum difference between two nearby mean values) then the two levels are considered separate classes. Thus, the boundary of classes is set according to the half of mean difference values. In the given example, the upper boundary of class 1 will be \((m1+(m2-m1)/2)\), which will be equal to the lower boundary of the second class if the mean values are in increasing order. The boundary value will be the same in case if the mean value is in decreasing order (i.e., \((m1-(m1-m2)/2) = (m1+ (m2-m1)/2)\)). In our experiments, the fourteenth feature, i.e., the font size, decreases (in general) with the increase in level, whereas in other features, mean values generally increase with the level of books.

• The multi-level classification model is built using the training data set. The threshold and other classes having different number of levels are classified according to the training dataset.

• Once the multi-level classification model is built, we test the accuracy of the test book by assigning it a different class for each feature. Within a class, the book is labeled with different possible levels (it can vary from 1 to many).

• After getting a set of possible levels (in one class) for each feature value of the test book, each level gets different scores depending on the labels assigned.
We have classified our training dataset as follows:

Graphs using the mean values per level of each feature for all the training set of 97 books:

Figure 15: Number of words in entire book versus level of the book
C1 = L[1,2,3]
C2 = L [4,5,6]
C3 = L [7,8]
C4 = L [9]
C5 = L [10]
C6 = L [11,12,13]
C7 = L [14]
Figure 16: Number of letters per word versus level of the book

C1 = L[1]
C2 = [2,3,4]
C3 = [5,6,7,8,9,10,11,12]
C4 = [13,14]
Figure 17: Number of sentences in entire book versus level of the book

C1 = [1,2,3,4]
C2 = [5,6,7,8,9]
C3 = [10]
C4 = [11,12,13,14]
Figure 18: Number of words per sentence versus level of the book

C1 = L[1,2,3,4,5,6,7,8,9,10,11,12,13]

C2 = L[14]
We discarded feature 5 because of its completely zigzag mean values and its inability to have continuous labels (levels) within a class.
Figure 20: Number of pages versus level of the book

C1 = L[1,2,3]
C2 = L[4,5,6]
C3 = L[7,8,9,10,11,12,13,14]
Figure 21: Number of words per page versus level of the book

\[ C_1 = L[1,2] \]
\[ C_2 = L[3,4] \]
\[ C_3 = L[5,6,7] \]
\[ C_4 = L[8,9] \]
\[ C_5 = L[10] \]
\[ C_6 = L[11] \]
\[ C_7 = L[12] \]
C8 = L[13]
C9 = [14]
Figure 22: Number of sentences per page versus level of the book

C1 = L[1,2,3,4]
C2 = L[5,6,7,8,9]
C3 = L[10]
C4 = L[11,12,13,14]
Figure 23: Number of textlines per page versus level of the book

C1 = L[1,2,3,4,5,6,7]

C2 = L[8,9,10,11,12,13]

C3 = L[14]
Figure 24: Number of words per textline versus level of the book

C1 = L[1,2,3,4,5,6,7,8,9]  
C2 = L[10]  
C3 = L[11,12,13,14]
Figure 25: Number of words per textbox versus level of the book

C1 = L [1,2,3]
C2 = L [4,5,6]
C3 = L [7]
C4 = L [8,9]
C5 = L [10]
C6 = L [11]
C7 = L [12,13,14]
C1 = L[1,2]
C2 = L[3,4,5,6,7]
C3 = L[8,9,10]
C4 = L[11]
C5 = L[12,13,14]
Figure 27: Average ratio of textbox area to page area versus level of the book

C1 = L[1,2]
C2 = L[3,4,5,6,7]
C3 = L[8,9,10,11,12,13]
C4 = L[14]
Figure 28: Average font size versus level of the book

\[ C_1 = L[1] \]
\[ C_2 = L[2,3,4,5,6,7,8,9] \]
\[ C_3 = L[10,11,12,13,14] \]
We will explain this methodology using an example from our experimental dataset.

Say we are considering the first book as our test book. The feature vector corresponding to first book is:

\[
[27.000, 3.370, 9.000, 3.000, 0.444, 4.000, 6.750, 2.250, 1.750, 3.857, 3.857,
1.000, 0.048, 27.391]
\]

Now, as shown in figure 15, first feature value (27.000) is classified under class $C_1$ where $C_1 = L[1,2,3]$.

Similarly, using other figures we can find that,

For the 2nd feature, it gets classified under $C_2 = L[2,3,4]$

3rd feature: $C_3 = L[1,2,3,4]$

4th feature: $C_4 = L[1,2,3,4,5,6,7,8,9,10,11,12,13]$

5th feature: We are not using this feature in our experiments because of its highly zigzag graph.

6th feature: $C_6 = L[1,2,3]$

7th feature: $C_7 = L[1,2]$

8th feature: $C_8 = L[1,2,3,4]$

9th feature: $C_9 = L[1,2,3,4,5,6,7]$

10th feature: $C_{10} = L[1,2,3,4,5,6,7,8,9]$

11th feature: $C_{11} = L[1,2,3]$

12th feature: $C_{12} = L[1,2]$

13th feature: $C_{13} = L[1,2]$
14th feature: C14 = L[1]

For computing (predicting) the level of each book, we calculate the scores for each level based on the class it has been assigned for each feature, evenly dividing a score of 1 for each feature across the levels assigned to a class. For the book described above, we can calculate $S(x)$ for each level $x$ corresponding to its level.

\[
S(1) = \frac{1}{3} + \frac{1}{4} + \frac{1}{13} + \frac{1}{2} + \frac{1}{4} + \frac{1}{7} + \frac{1}{9} + \frac{1}{3} + \frac{1}{2} + \frac{1}{2} + 1 = 4.33
\]

\[
S(2) = \frac{1}{3} + \frac{1}{3} + \frac{1}{4} + \frac{1}{13} + \frac{1}{3} + \frac{1}{2} + \frac{1}{4} + \frac{1}{7} + \frac{1}{9} + \frac{1}{3} + \frac{1}{2} + \frac{1}{2} = 3.66
\]

\[
S(3) = \frac{1}{3} + \frac{1}{3} + \frac{1}{4} + \frac{1}{13} + \frac{1}{3} + \frac{1}{4} + \frac{1}{7} + \frac{1}{9} + \frac{1}{3} = 2.16
\]

\[
S(4) = \frac{1}{3} + \frac{1}{4} + \frac{1}{13} + \frac{1}{4} + \frac{1}{7} + \frac{1}{9} = 1.16
\]

\[
S(5) = \frac{1}{13} + \frac{1}{7} + \frac{1}{9} = 0.33
\]

\[
S(6) = \frac{1}{13} + \frac{1}{7} + \frac{1}{9} = 0.33
\]

\[
S(7) = \frac{1}{13} + \frac{1}{7} + \frac{1}{9} = 0.33
\]

\[
S(8) = \frac{1}{13} + \frac{1}{9} = 0.19
\]

\[
S(9) = \frac{1}{13} + \frac{1}{9} = 0.19
\]

\[
S(10) = S(11) = S(12) = S(13) = \frac{1}{13} = 0.07
\]

\[
S(14) = 0
\]

Based on the above scores for each level, we predict the reading level by averaging the levels corresponding to the top three scores. Doing an average gives better prediction when we get almost similar scores for more than one
feature. We calculate the accuracy, claiming it is correct if the predicted book level is within +/- 1 of the true reading level.
4. Substantiation of the readability assessment framework using different approaches

4.1. Performance results using SVM-ranking

For training data: With a training data set of 97 books, we performed the forward–feature optimization algorithm followed by our modified SVM-rank algorithm for the readability assessment framework to determine the prediction accuracy (performance). We have used two metrics as used in Ma et al. (2012a):

1. Accuracy: Accuracy of the readability assessment framework is calculated by claiming it is correct if the predicted book level is within +/- 1 of the true reading level.

2. Average leveling error: It is the absolute error of number of levels away from the key reading level, averaged over all of the books.
Results using single feature:

<table>
<thead>
<tr>
<th>Features</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.55</td>
<td>0.38</td>
<td>0.49</td>
<td>0.29</td>
<td>0.39</td>
<td>0.46</td>
<td>0.37</td>
<td>0.36</td>
<td>0.38</td>
<td>0.55</td>
<td>0.53</td>
<td>0.37</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>Average leveling error</td>
<td>1.44</td>
<td>2.59</td>
<td>1.88</td>
<td>2.86</td>
<td>3.42</td>
<td>2.48</td>
<td>1.89</td>
<td>2.35</td>
<td>2.12</td>
<td>2.3</td>
<td>1.59</td>
<td>1.67</td>
<td>2.45</td>
<td>2.25</td>
</tr>
</tbody>
</table>

Table 1: Prediction accuracy with single feature

We have taken the best 5 features, using their performance as shown in the above table: i.e., 1,3,7,11,12.

After selecting the best five features, we performed the forward–feature optimization algorithm with the addition of a single best feature at each step. The following tables show the accuracy and average leveling error at each step of the algorithm:
Table 2: Prediction accuracy for the best five features i.e., feature 1,3,7,11,12 and with the addition of one single feature at a time from the rest of the features

Next set of best features: 1,3,5,7,11,12 (Point to consider: Although feature 5, when analyzed alone, has given the worst performance, it has given the best performance to this point).

Table 3: Prediction accuracy of the six best features with an addition of one feature from the remaining subset of features
Table 4: Prediction accuracy of the seven best features with the addition of one feature from the remaining subset of features

<table>
<thead>
<tr>
<th>Features</th>
<th>1,2,3,5,7,11,12</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.68</td>
<td>0.71</td>
<td><strong>0.72</strong></td>
<td>0.69</td>
<td>0.67</td>
<td>0.67</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>Average leveling error</td>
<td>1.13</td>
<td>1.86</td>
<td>1.71</td>
<td>1.26</td>
<td>1.37</td>
<td>1.29</td>
<td>1.09</td>
<td>1.69</td>
</tr>
</tbody>
</table>

Table 5: Prediction accuracy of the eight best features with the addition of one feature from the remaining subset of features

<table>
<thead>
<tr>
<th>Features</th>
<th>1,2,3,5,6,7,11,12</th>
<th>4</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.72</td>
<td>0.71</td>
<td>0.72</td>
<td>0.73</td>
<td><strong>0.79</strong></td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>Average leveling error</td>
<td>1.71</td>
<td>1.09</td>
<td>1.02</td>
<td>1.08</td>
<td>1.0</td>
<td>1.07</td>
<td>1.02</td>
</tr>
</tbody>
</table>

Table 6: Prediction accuracy of the nine best features with the addition of one feature from the remaining subset of features

<table>
<thead>
<tr>
<th>Features</th>
<th>1,2,3,5,6,7,10,11,12</th>
<th>4</th>
<th>8</th>
<th>9</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.79</td>
<td>0.77</td>
<td><strong>0.81</strong></td>
<td>0.78</td>
<td>0.76</td>
<td>0.73</td>
</tr>
<tr>
<td>Average leveling error</td>
<td>1.0</td>
<td>0.98</td>
<td>0.92</td>
<td>1.01</td>
<td>0.98</td>
<td>1.04</td>
</tr>
<tr>
<td>Features</td>
<td>1,2,3,5,6,7,8,10,11,12</td>
<td>4</td>
<td>9</td>
<td>13</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td></td>
<td>0.81</td>
<td>0.81</td>
<td>0.80</td>
<td><strong>0.82</strong></td>
<td>0.79</td>
</tr>
<tr>
<td><strong>Average leveling error</strong></td>
<td></td>
<td>0.92</td>
<td>0.96</td>
<td>0.96</td>
<td>0.94</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 7: Prediction accuracy of the ten best features with the addition of one feature from the remaining subset of features

<table>
<thead>
<tr>
<th>Features</th>
<th>1,2,3,5,6,7,8,10,11,12,13</th>
<th>4</th>
<th>9</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td></td>
<td>0.82</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Average leveling error</strong></td>
<td></td>
<td>0.94</td>
<td>1.03</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table 8: Prediction accuracy of the eleven best features with the addition of one feature from the remaining subset of features

As can be seen in Table 8, there is no further improvement from adding another feature. According to our algorithm, we discontinue further addition of features from now onwards. Hence, the best set of features using training data: 1,2,3,5,6,7,8,10,11,12,13 gives an accuracy of 82% as compared to 76% of accuracy obtained using all 14 features.
Table 9: Prediction accuracy of all 14 features with training dataset

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
<th>Average leveling error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2,3,4,5,6,7,8,9,10,11,12,13,14</td>
<td>0.76</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Table 10: Results on evaluation data set with completely new data set of 38 books

<table>
<thead>
<tr>
<th>Features</th>
<th>Accuracy</th>
<th>Average leveling error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,3,7,11,12</td>
<td>0.66</td>
<td>1.19</td>
</tr>
<tr>
<td>1,3,5,7,11,12</td>
<td>0.61</td>
<td>1.31</td>
</tr>
<tr>
<td><strong>1,2,3,5,7,11,12</strong></td>
<td>0.58</td>
<td>1.37</td>
</tr>
<tr>
<td>1,2,3,5,6,7,11,12</td>
<td>0.73</td>
<td>1.15</td>
</tr>
<tr>
<td>1,2,3,5,6,7,10,11,12</td>
<td>0.68</td>
<td>1.21</td>
</tr>
<tr>
<td>1,2,3,5,6,7,8,10,11,12</td>
<td>0.68</td>
<td>1.13</td>
</tr>
<tr>
<td>1,2,3,5,6,7,8,10,11,12,13</td>
<td>0.66</td>
<td>1.15</td>
</tr>
<tr>
<td>All 14 features</td>
<td>0.61</td>
<td>1.39</td>
</tr>
</tbody>
</table>
With the evaluation dataset, we got 66% accuracy using the features 1,2,3,5,6,7,8,9,10,11,12,13, whereas we got 61% accuracy with all fourteen features.

The performance results after feature selection (82% versus 76% accuracy in the training dataset; 66% versus 61% accuracy in the evaluation dataset) are not statistically significant, but still this analysis has thrown light on some of the features’ behaviors. The best performance is observed with features 1,2,3,5,6,7,8,10,11,12,13 in the training data set.

The features which got eliminated during feature selection are:

- Feature 4: Number of words per sentence.
- Feature 9: Number of textlines per page.
- Feature 14: Average font size

These features have mostly reduced the performance of the system whenever they got added. As can be seen in Table 1, these features are among the ones which have performed worst. This can actually lead to a selection methodology where we can do bi-directional feature selection, i.e., both addition of the best performing feature and elimination of the worst performing feature (the one whose addition actually reduces the performance as compared to the feature set without that feature). This can give a fast feature selection methodology and can be considered important when there is a large number of features or when the
learning algorithm is expensive to perform. We could have gotten the same result if we had followed this bi-directional feature selection approach.

The worst performance of feature 4, i.e., number of words per sentence, can be attributed to our automatic annotation process. As can be seen in Figure 4, the data values are quite wide-spread, and there are quite large numbers of outliers. During data preparation, a lot of errors got introduced while generating OCR from the pdf book. The most common error, which has not much effect on other features but has affected the “number of words per sentence,” is the addition of “.”(dot). This error got introduced mostly from the following two factors:

- misread symbols (when OCR misreads text and symbols)
- false alarms (when OCR recognizes pictures as some text; mostly it recognizes them as some combination of “.”).

These extra dots are read as “full stop” and, hence, generate more sentences than the actual. These OCR errors are quite dependent on the quality of the book printing (and some more external factors) and not really dependent on the level of book. Heuristic data checking of the OCR output could have been used to clean up some of the data, but it was not done for this study.

Feature 14 (Average font size) has also not given good performance. As can be seen from figure 14, there is not much difference in font size across levels. Probably it can be considered on a much coarser level (for example, grade 1 and grade 4, etc.).
After getting the results from the evaluation dataset, we have found that this greedy approach has not given the best performance but, still, a more improved performance than having all the features. In the evaluation set, we have gotten the better performance with 1,2,3,5,6,7,11,12 features. This approach might not give an optimal solution (as suspected!).
4.2. Results using multi-level classification

In the multi-level classification algorithm, we have used the following metric to calculate the system performance:

1. Accuracy: Accuracy of readability assessment framework is calculated by claiming it is correct if the predicted book level is within +/- 1 of the true reading level.

The prediction accuracy with 97 books of the training dataset is 61%, and the prediction performance using 38 new books is 66%.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training dataset (97 books)</td>
<td>61%</td>
</tr>
<tr>
<td>Evaluation dataset (38 books)</td>
<td>66%</td>
</tr>
</tbody>
</table>

Table 11: Results using multi-level classification

This method has not given a better performance as compared to results we have obtained in our first set of experiments. But we have quite an equivalent performance in our evaluation dataset.

We were expecting to get reduced accuracy in the evaluation dataset as compared to the training dataset because of some potential overfitting of data. But in spite of our speculation, the performance was better in the evaluation dataset as compared
to the training dataset. The performance in the training dataset could be low in comparison to the evaluation dataset because of the large dataset (97 books as compared to 38 books), which increases the error probability (increased probability of outliers).

We have not used feature 5, i.e., ‘type-token ratio’, in the feature vector because we are not able to classify its data into classes, as there is too much variance from level to level.

**Comparative analysis of multi-level classification versus SVM-ranking**

The smaller number of possible classes in features 4 (Number of words per sentence, figure 18), 9 (Number of textlines per page, figure 23), 14 (Average font size, figure 28) depicts the high variance of data from level to level, and this variance is also affirmed with the poor performance of these features during the feature selection process (Table 1) in SVM-ranking based experiments. The greater the number of classes in features (in multi-level classification experiments), the less the variance in data; and this is coherent with the better performance of individual features during the feature selection process (Table 1). Hence, we can conclude that multi-level classification has quite a good predictability of relevant features when compared with the SVM-based feature selection algorithm. The time and resources required for the SVM-based feature selection method is quite high as compared to the multi-level classification method.
The ‘type-token ratio’ feature was excluded in multi-level classification experiments. This feature has performed worst in SVM rank experiments and also when it was considered independently (Table1). But in later experiments of feature selection, this feature gets included in the final set of features selected. When we performed experiments with the evaluation dataset (Table 9), this feature has actually reduced the performance of the overall system when it gets added.

We also suspect that the performance of the multi-level classification method can be improved if we incorporate an even more finely tuned property of data, which can consider the variance of data more appropriately (as compared to mean, which generalizes the results).

We expect to have a more normalized feature space if we have an equal number of books per level and from the same publishers. According to our analysis of data, we have seen that some publishers have a quite different layout of books than others. For example, some publishers have lower grade books which have quite high feature values (almost equal to 3-4 levels higher). The feature total number of words in the entire book (feature 1) has given very good prediction results. These features are quite dependent of the organization of the book, although our feature set should be independent of these factors. But if this analysis would have been done for one publisher, then we could have gotten better performance.
5. Conclusion and Remarks

In this study, we did analysis of various features we have used in our readability assessment framework designed for elementary children’s literature. We did feature selection to understand the nature of features which are most relevant in enhancing the performance of the framework. We have found some of the features which hampered the predicting performance. Although we are able to increase the performance, the results are not statistically significant. With our simple multi-level classification algorithm, we are able to get a good performance. Overall, we are able to find the features which contributed the most to the performance of the framework.

In the future, we are interested in automating the process of extracting more visually-oriented features. It has been proven (Ma et al., 2012a) that the visually oriented features give a very significant contribution to the predicting performance. We have incorporated some of the features in our study. We have not incorporated the features related to image layout, for example, the ratio of image area to textbox area, etc. If we can get more image-oriented features, then we are hoping to get even more improved performance of our framework.
References


J. S. Chall. 1958. Readability: An appraisal of research and application. number-34. Ohio State University