IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 4 May 2023, accepted 18 May 2023, date of publication 24 May 2023, date of current version 7 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3279360

RESEARCH ARTICLE

Comprehensive Readability Assessment of Scientific Learning Resources

MUDDASSIRA ARSHAD[1,2], (Graduate Student Member, IEEE), MUHAMMAD MURTAZA YOUSAF[2], AND SYED MANSOOR SARWAR[3], (Member, IEEE)

[1]Department of Computer Science, University of the Punjab, Lahore 54000, Pakistan
[2]Department of Software Engineering, University of the Punjab, Lahore 54000, Pakistan
[3]University of Engineering and Technology, Lahore 54890, Pakistan

Corresponding author: Muddassira Arshad (muddassira@pucit.edu.pk)

ABSTRACT Readability is the measure of how easier a piece of text is. Readability assessment plays a crucial role in facilitating content writers and proofreaders to receive guidance about how easy or difficult a piece of text is. In literature, classical readability, lexical measures, and deep learning based model have been proposed to assess the text readability. However, readability assessment using machine and deep learning is a data-intensive task, which requires a reasonable-sized dataset for accurate assessment. While several datasets, readability indices (RI) and assessment models have been proposed for military agencies manuals, health documents, and early educational materials, studies related to the readability assessment of computer science literature are limited. To address this gap, we have contributed Computer science (CS) literature dataset **AGREE**, comprising 42,850 learning resources(LR). We assessed the readability of learning objects(LOs) pertaining to domains of Computer Science (CS), machine learning (ML), software engineering (SE), and natural language processing (NLP). LOs consists of research papers, lecture notes and Wikipedia content of topics list of learning repositories for CS, NLP, SE and ML in English Language. From the statistically significant sample of LOs two annotators manually annotated LO's text difficulty and established gold standard. Text readability was computed using 14 readability Indices (RI) and 12 lexical measures (LM). RI were ensembled, and readability measures were used to train the model for readability assessment. The results indicate that the extra tree classifier performs well on the AGREE dataset, exhibiting high accuracy, F1 score, and efficiency. We observed that there is no consensus among readability measures for shorter texts, but as the length of the text increases, the accuracy improves. The AGREE and SELRD datasets, along with the associated readability measures, provide a novel contribution to the field. They can be used to train deep learning models for readability assessment, develop recommender systems, and assist in curriculum planning within the domain of Computer Science. In the future, we plan to scale AGREE by adding more LOs and adding multimedia LOs. In addition, we would explore the use of deep learning methods for improved readability assessment.

INDEX TERMS Automated readability index, CS learning resource repository, Flesch Kincaid reading ease, Flesch Kincaid grade index, gunning fog readability index, lexical diversity, lexical richness, lexical chains, Lix, new Dale-Chall, readability assessment, readability gold standard for CS learning objects.

I. INTRODUCTION

In the context of COVID and post-COVID economic crises, self-learners are highly motivated to improve their skill set and gain a better understanding of various subjects [1], [2]. However, they may face several problems in accessing learning resources, such as matching the online content with their learning level and time constraints [3]. Readability assessment plays a vital role in addressing these concerns by guiding learners in selecting appropriate resources that match their reading level and minimize digital inequality. The significance of ability to assess the readability of a piece

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés.

of text for content writers and proofreaders, can aid in text simplification, resource recommendation, and curriculum planning.

Plavén-Sigray et al. [4] in his study conducted on research papers pertaining to the field of Life Sciences, concluded that readability of the scientific literature is decreasing over time. In addition, Hayes [5] carried out the study to assess readability of the technical content from ten professional journals in astronomy, biology, chemistry, geology and physics; in addition to the science textbooks for introductory courses offered at college level, and the popular science magazines. Hayes concluded that scientific content comprehension has become increasingly challenging for individuals without specialized knowledge. Although, these studies were conducted on diversified subjects, however, learning resources from CS domain were not included.

Readability is defined as the quality of being legible and decipherable. Readability Indices (RIs) are based on objectively measurable and dependable text characteristics [6]. These characteristics could be as atomic as tokens count, sentence length [7], [8] and may vary to include part-of-speech tags, phrase level information. The readability of the content can be assessed using lexical measures, readability indices (RI), or state-of-the-art transformer-based models [9], [10], [11].However, such approaches are data intensive and require specialized and large datasets.

The datasets being used for evaluation of general readability assessments are Weebit [12], OneStopEng [13], Cambridge [14] and CommonLit Ease of Readability(CLEAR) [15], where OneStopEng, CLEAR and Cambridge are publically accessible. These datasets either refer to the general articles being published [12], or comprises English passages for evaluating readability level of L2 learners [14]. Few datasets for financial system readability assessment [16] or software code readability [17] have also been discussed in literature to assess the readability in specific domains. These datasets and the earlier studies conducted on readability assessment of scientific literature are either generic and do not refer to CS domain learning resources. Therefore, there exist a research gap of readability assessment of Computer Science based Learning Resources (CSLR).

To fill this gap, and to address the challenges of automatic readability assessment mechanisms, we designed our dataset consisting of Software Engineering Learning Resources (SELRD) and aggregated it by unifying six existing datasets: Lecturebank [18], TutorialBank [19], ACL Anthology Network Corpus [20], AL-CPL [21], University Course Dataset and NPTEL MOOC [22] dataset to yield AGREE dataset.

The AGREE dataset would not only be used as a repository of CS learning resources in English Language, but may also be used for prerequisite chain learning, reading list generation, text simplification and survey extraction.

The purpose of the study was to investigate and address the following research inquiries:

RQ1: Is there a consensus among the existing readability measures for Computer Science LOs

RQ2: Does readability statistics matches with actual difficulty level/gold standard?

RQ3: Is there a consensus between reading time's and readability measures?

RQ4: Which classifier serves as a suitable machine learning model for our dataset to assess difficulty level on the basis of readability indices and lexical measures?

RQ5: Is there a Computer Science scientific literature data set to facilitate deep learning models for readability assessment?

In order to seek answers for these research questions, the comprehensive study was carried out. The preamble, methodology and findings of the research are organized as follows: Section II consists of related work, In Section III, we outline the measures and metrics utilized in our study. Section IV delves into the datasets employed in our study, while Section V focuses on the preprocessing steps applied to our dataset. The methodology utilized is discussed in Section VI, followed by the gold standard annotation process in Section VII. Section VIII synthesizes the results and initiates discussions, while Section IX concludes the research by summarizing key findings and outlining future work.

Our study contributes to the growing body of research on readability analysis of scientific literature, and provides valuable insights into the readability of CSLRs. The contribution of the Software Engineering Learning Resources Dataset (SELRD) and AGREE Readability Data set for Computer Science Literature with their associated readability measures, and reading time is novel and comprehensive approach to assess the text readability of CSLR.

II. RELATED WORK

Textual difficulty means how easy or hard a text is to read. Research has shown that both the content and presentation are main factors affecting the ease with which texts are read. The difficulty level of content has been assessed using vocabulary, averages of word lengths and sentences length. Moreover, number and types of syllables have also been used for several decades for readability assessment. Some approaches have also been developed which uses linguistic features [23] part-of-speech tags, lexical chain length, and lexical difficulty. Lexical Chain refers to sequence between the semantically related ordered words and thus represent the text cohesion [24]. The term lexical difficulty refers to the difficulty of words which are less common, are longer polysyllable words. Less common words are typically not as familiar to readers compared to common words. These rare and challenging words are often longer and may pose difficulties for readers in terms of comprehension. In the field of readability assessment, there has been a growing utilization of various machine learning and deep learning techniques [25], [26], and [11]. These advanced approaches have significantly contributed to improving the accuracy of suggesting appropriate readability levels for texts. By harnessing the power of machine learning and deep learning, readability assessment methods have been able to provide more precise and tailored readability recommendations, enhancing the overall reading experience for individuals. The absence of Computer Science-related datasets in the field of readability assessment is a significant challenge. These approaches heavily rely on data, particularly large datasets, for effective training and accurate predictions. While there are several general-purpose datasets and corpora available for readability assessment in the English language, such as WeeBit, OneStopEng, and CLEAR, the lack of Computer Science-related datasets limits the applicability and relevance of these approaches in the CS domain.

The importance of domain-specific datasets cannot be overlooked, as they play a crucial role in tailoring readability assessment techniques to specific domains such as Finance and Medical. These datasets provide valuable insights into the language and content characteristics of the respective domains, enabling more accurate and context-aware readability assessments.

However, it is worth noting that even within the existing datasets, some large-sized datasets relevant to readability assessment are in foreign languages like Chinese [27]. This further highlights the need for diverse and comprehensive datasets in multiple languages, including Computer Science-related content, to ensure the effectiveness and generalizability of readability assessment techniques across different domains and languages.

Crossley et al. [15] contributed CLEAR corpus of around 5000 text excerpts selected from open sources for 3rd to 12th graders from various genres, such as informational and literary texts. Trained teachers rated the readability of each excerpt based on student comprehension. The corpus was evaluated using ARI, NDC, FKEL, FKGR, Crowd Sourced Algorithm for Reading Comprehension (CAREC), and Cohesion Matrix L2 Readability Matrix (CML2RI). The text was also evaluated using lexico-semantics and higher-level semantics [28], Syntactic Sophistication and Complexity [29] measures. CLEAR includes a unique readability score (CLEAR score) for each excerpt, which allows for modeling individual texts instead of grouping them by difficulty level. The statistical analyses found that literary excerpts were easier to process than informational ones, and more recent excerpts were easier to understand than older ones due to language change over time.

OnestopEnglish(OEC) corpus is contributed for readability assessment providing the functionality of automatic readability assessment and text simplification. The OEC was created from the content published on onestopenglish.com over the period 2013 to 2016 by collecting parallel versions of texts at easy, intermediate, and advanced difficulty levels [13]. The readability of the corpus was evaluated using various measures, including character and word n-grams, POS, dependency relations, and traditional readability assessment and lexical measures. The Sequential Minimal Optimization (SMO) algorithm was used to train a classifier, and the results showed that the highest classification accuracy of 78.13 percent was achieved when all the linguistic complexity features, such as syntactic, discourse-based, psycho-linguistic, and traditional measures were included. Even though the corpus is available for free, it has limited size and does not pertain to any specific domain.

Zhou et al. [30] conducted the comparative study to evaluate five readability indices for technical material instead of general content. These readability indices include i.e. Flesch Kincaid Grade Reading Level(FKGR) [31], Gunning Fog (GFRI), Simple Measure of Gobbledygook (SMOG), Coleman Liau (CLRI), and Automated Readability Indices(ARI) [32]. Zhou et al. [30] used the readability assessment automated tool available in Microsoft Word, and websites: readability-score.com,[1] readable.com,[2] onlineutility.org and edicentral.com. These tools were shortlisted from Google Search. While MS Word only provide readability statistics using Flesch Kincaid grade level (FKGR),and Flesch Kincaid reading ease (FKEL), Readability-score.com and Readable.com provided greater flexibility in terms of providing input prompt for adding text directly in addition to provision of providing URL to refer the web resource. It specifies readability using FKEL, GFRI, SMOG, ARI, CLRI in addition to presenting the statistics indicating count of sentences, words, complex words, and averages of words per sentences and syllables per word. In addition, it also specifies percentage of complex words. Edicentral support was not available. Further more, readable.com provides support to FORCAST Grade Level (FGL), Lasbarhets index Swedish Readability Formula(LIX) and RIX, Fry Readability Index (FRI), Raygor Readability Index (RRI), Common European Framework of Reference for Languages(CEFR), Spache Score (SSRI), Linsear Write (LRRI), New Dale Chall Score (NDC), IELTS Level and Powers Sumner Kearl Grade (PSKG) in addition to other simple approaches like FKEL, GFRI, SMOG, ARI and CLRI.

Several other web resources to assess the readability formula are also available like Readability formulas[3] computed New Dale-Chall(NDC), FKEL, FKGR, FRI, GFRI, PSKG, SMOG, FGL, and SSRI. The study concludes that most of these formula asses the word count, syllables, and/or sentence count which to some extent can assess the text readability. In addition, it was found that various measures converge as the length of the paragraph increases upto 900words. The differences occur due to acronyms, hyphenated words, and punctuation. It was also pertinent to note that the organization of the content, the overall presentation using text formatting [33], and most of the punctuation marks are not considered in these formulae.

Bormuth [7] recommends the use of cloze test to understand the text difficulty level. He concluded to use nonlinear correlation for effective readability formula. Moreover, the scope of the readability could be extended to word, clauses or sentences level.

[1] https://www.webfx.com/tools/read-able/

[2] http://www.readable.com

[3] https://readabilityformulas.com

Boudjella et al. [34] discusses the comparative analysis of students between grade level 9 and 14. Students' average grade level was assessed using readability models. LOs were presented using font size 12 and 14. Students were asked to read the text, reading time was recorded. The student's comprehension was assessed using the MCQs based questioner. Linear regression was used to model the results. It was concluded that reading speed is directly proportional to the score. When the reading speed decreases, score also decreases. In addition, reading speed of native learners (L1) is better as compared to non-native (L2) speakers.

Brantmeier [35] discusses the use of ANOVA, Regression Models, MANOVA for analysis. In order to assess readability, it was concluded that ANOVA test has high appropriateness for testing inferences. Feng [36] in her Doctoral thesis included features including Lexical chains length and span, count of lexical chains, in addition to the count and average of entity mentions and unique entities. The proposed approach improved the accuracy by 70% as compared to Flesch Kincaid grade level. However, it has increased complexity due to multiple measures.

Kiselnikov et al. [37] presented Coh-metrix for academic texts with consideration on semantics, morphology instead of sentence and word structures. Coh-metrix readability is computed as a weighted sum of logarithmic mean of words (CELEX word frequency), sentence to sentence adjacent mean(Sentence Syntax Similarity) and proportional unweighted adjacent sentences (Content Word Overlap). The correlation between FKEL and Coh-metrix is 0.626. However, Coh-metrix has high complexity in contrast to FKEL. In addition, Coh-metrix tool support[4] allows text size of 15000 characters, without any special characters, which was not applicable to our dataset. The AAN dataset has an average character count of around 29,000 characters, while AL-CPL has 60,000 characters. NPTEL consists of approximately 28,000 characters, LB has 6,900 characters, TB has 6,400 characters, and UCD has 710 characters, including special characters, per learning object (LO). Since AAN, AL-CPL and NPTEL did not meet Coh-Metrix constraints, we have not incorporated it in our work.

Plavén-Sigray et al. [4] in their study of readability assessment of scientific literature used the FKEL and NDC, and concluded that the readability of scientific text is decreasing over time. The study was based on abstracts of 709,577 research articles published in 123 journals over a span of 130 years (1881 to 2015). On the basis of readability analysis, the study emphasizes reducing the scientific jargon to improve readability and accessibility.

Xia [38] presented lexicosemantic, parse-tree, language modeling, discourse based features to assess readability of text for native and second language (L2) learners. Their study concluded that classification model for readability assessment outperforms regression models.

Lix [39] readability index measures the difficulty of a foreign language using words, word lengths and punctuation like period, colon or capitalized text. This is an efficient mechanism and an easy to calculate solution as it does not include syllables or poly syllables.

These readability formula have multiple applications like inter subject difficulty predictability, and valid prediction of difficulty level of clauses, sentences, and words. In addition, the approaches based on linguistic variables may improve results validity. Therefore, we conclude that several datasets and techniques have been proposed. These datasets are either based on general purpose text being classified as fiction or informative text, does not contain large corpus size or are not publicly available. In addition, readability assessment techniques range from simple counts of syllables to lexical analysis and machine learning approaches to deep learning approaches. We intend to apply the readability and lexical measures in addition to machine learning measures to large scale learning repository of CS, ML, and NLP. To our knowledge, it is the first attempt to establish comprehensive readability assessment for CSLR. Our scope includes designing a large scale CSLR covering LO as well as their readability indices for readability assessment on the basis of text content extracted from LOs and not the content's formatting/legibility.

III. METRICS AND MEASURES USED

In literature, several readability formula have been used to gauge the readability of the text. These formulae are based on text statistics. These statistics vary from letter count to syllable and token counts. We have computed the following measures presented in tables [1],[6] on statistically significant sample extracted of size 1448 LR from AGREE dataset. These measures were helpful in computing readability indices and their averages. We have excluded the metrics Spache, Power-Sumner-Kear, TextEvaluator [40] as these measures are recommended to assess readability for primary grade students. We did not use Coh-Metrix [37] as it is strongly correlated with easier, simpler and efficient mechanisms of Flesch Kincaid Grade Level and reading ease [41].

A. DIFFICULTY LEVEL BASED INDICES

In literature few indices have been proposed to assess the text difficulty level. These indices present the qualitative measure of text difficulty. Difficulty based RI include Flesch Kincaid Reading Ease and Mcalpine Eflaw score.

1) MCALPINE EFLAW SCORE

Although one of the major problems for L2 learners is understanding the longer sentences, but cluster of miniwords also poses problem in understanding. The readability is assessed in terms of the 'flaw' in the text which is computed in terms of miniwords frequency. Therefore, McalpineEflaw Score[5]

[4] http://tool.cohmetrix.com/

[5] https://strainindex.wordpress.com/2009/04/30/mcalpine-eflaw-readability-score/

TABLE 1. Readability measures used.

Measure	Abbr	Library	Comments
CharacterCount	CC	Self	Spaces ignored
Character Per Word	CPW	Self	CC to WC ratio
Complex Word Count	CWC	Self	Word not in Dale-Chall list
Letter Count	LeC	Self	Punctuation Off
Lexicon Count	LC	Self	punctuation off
Longword Count	LWC	Self	WC with length > 6
MonoSyllable Count	MSC	Self	Single Syllable words
PolySyllable Count	PSC	Self	> 3 syllables
Paragraphs Count	PC	Self	Paragraph count
Sentence Count	SC	Pyphen	Total sentences in LO
Sentences per Paragraph	SPP	Self	SC to PC ratio
Syllable Count	SyC	Pyphen	PSCs + MSC
Type per Token	TPT	Self	unique words count/WC
Word Count	WC	Self	Count of words in LO
Word per Sentences	WPS	Self	LC to SC ratio

TABLE 2. Eflaw score interpretation.

Score	Interpretation
1 -20	very Easy
21 - 25	Quite Easy
26-29	Little Difficult
> 30	Very Confusing

was suggested. Considering the count of words (W), Mini words(M) defined as the words comprising on 1, 2 or 3 letters and sentences (S),

$$McAlpineEFlaw = \frac{W + M}{S} \quad (1)$$

Table [2] refers to of Mcalpine Eflaw score's (1) interpretation.

Table [2] suggest that text scoring > 30 is very difficult and therefore for universal readability, Eflaw score up to 29 is acceptable.

2) FLESCH KINCAID READING EASE

With the motivation to facilitate the recommendation for reading ease, FKEL specifies that if the number of syllables in a word increases, so does its difficulty to read and understand. Eq [2] computes the FKEL.

$$FKEL = 206.635 - 1.105(\frac{WC}{SC}) - 84.6(\frac{SyC}{WC}) \quad (2)$$

where WC is word count, SC is sentence count, SyC is syllables count The result of (2): are interpreted to consider 60-69 as a universal reading ease. Scores below 10 are extremely difficult whereas scores above 90 are extremely easy. (2) shows that text readability is directly proportional to syllables counts. In addition, FKEL is directly proportional to reading ease.

3) LINSEAR WRITE(LRRI)

Linsear write is used to compute readability of the text in terms of weighted average of monosyllables and polysyllables in the text. For this purpose, 100 words sample is extracted from the text, from which monosyllables and polysyllables are processed, and their weighted average is computed as a raw score. The Raw score is processed to yield US grade level.

TABLE 3. FKGR interpretation.

Score	Interpretation
0-6	Beginner/very easy
6- 10	Average
>10	Skilled

$$RawScore = \frac{MSC + (3PSC)}{SC}$$

$$LRRI = \begin{cases} \frac{RawScore - 2}{2}, & \text{if} RawScore \leq 20 \\ \frac{RawScore}{2}, & \text{if} otherwise \end{cases} \quad (3)$$

Here MSC represent Monosyllable count, PSC: Polysyllable count, SC: sentence count. A universally readable text is typically characterized by an LRRI value ranging from 70 to 85. Texts with an LRRI value greater than 85 are considered difficult and challenging to read. On the other hand, texts with an LRRI value below 70 are categorized as very easy and straightforward to read. These LRRI thresholds provide a general guideline for assessing the readability levels of texts based on their LRRI values. However, LRRI value depends upon the 100 words sample selected. Heuristics may be devised so that multiple 100 word samples can be processed and their results can be ensembled.

B. US GRADE LEVEL INDICES

Using the content metrics like counts and averages of terms, words, sentences, and syllables, LOs difficulty for US grade level students can be established. Following metrics presents direct recommendation of US grade level. However, these indices do not reflect difficulty level for L2 learners.

1) FLESCH KINCAID GRADE LEVEL

The value of Eq.[2] refers to the reading ease, however, FKGR refers the recommended US grade. For universal readability, US grade level of 8 is suggested.

$$FKGR = (0.39)(ASL) + (11.8)(ASyW) - 15.59 \quad (4)$$

where ASL: average sentence length, ASyW: average number of syllables per word. The value of Eq [4] specifies US grade level's intellect to understand the content. Table [3] interprets the results of (4)

2) SMOG

G. Harry McLaughlin maps the difficulty of LOs to the pollution concept in text, stating that the difficulty of the text is actually a problem within the text With the constraint of having minimum 30 sentences to present reliable results, the results have been widely adopted for providing accessibility in health resources.

$$SMOG = 3 + Round(\sqrt{PSC}). \quad (5)$$

PSC is the count of Polysyllable words Eq (5).

The result of square root operation is rounded to nearest 10. The result of (5) specifies US grade level education required to understand the text under consideration. Pearson Correlation Coefficient between comprehension and SMOG is 0.88, showing that actual reading comprehension correlates with the results produced by SMOG.

3) AUTOMATED READABILITY INDEX (ARI)

In order to provide automated counting method for the text, syllables count may be time-consuming. Therefore, it was proposed to have simpler and efficient mechanism. For this purpose, characters instead of syllables were used for counting. This gives results similar to other frameworks with more efficiency. ARI is suitable for assessing technical documents.

$$ARI = 4.71 + (\frac{CC}{WC}) + 0.5(\frac{WC}{SC}) - 21.43 \tag{6}$$

where CC represents Character Count, WC: Word Counts, SC: Sentences Count.

4) GUNNING FOG (GFRI)

Fog index [42] is also considered to be one of the reliable mechanisms to assess the text difficulty level. Based on ratios of words to sentences count, and complex words to words, the metric calculates suitability of the text by recommending US grade level. Complexity is defined in terms of words with at least three syllables. However, since the words are assessed for the complexity, it is less efficient then formula based on simple words or sentences count.

$$GFRI = 0.4((\frac{WC}{SC}) + 100(\frac{CWC}{WC})) \tag{7}$$

The value of Eq[7] specifies US grade level required to understand the given text. Universal readability is attained for GFRI value lies between 7.0 to 8.0.

5) COLEMAN LIAU (CLRI)

In order to assess the readability of the textbooks, a formula based on computerized assessment of text for sentence(SC), characters(CC) and word (WC) lengths was established using OCR. Extracted lengths were averaged per 100 words. Considering L as average character counts per 100 words, and S as average sentences per 100 words, Coleman Liau readability index CLRI is computed as:

$$CLRI = ((0.0588)(L)) - ((0.296)(S) \; - 15.8 \tag{8}$$

The result of Eq.8 indicates the US grade level for which this text is easily understandable. For the universal readability, US grade level 8 - 10 is recommended. In order to improve accuracy, the researchers prefer word count $\geq$ 300. CLRI has proven its usage in medical, law as well as education sector.

6) NEW DALE-CHALL (NDC)

The New Dale-Chall indicator is based on list consisting of 3000 words being considered to be suitable for a US fourth grade student. New Dale-Chall list has improved vocabulary, plurality and tenses addition as compared to original proposed version of 763 words. The formula is based on the words which are not in the list as well as average length of the sentence. In Equation [9], RawNDC (RNDC) score is adjusted only when the ratio of difficulty words in the text increases by 5%. The adjusted score (NDC) is later interpreted using a predefined scale to get US Grade Level suitability. NDC is interpreted using table [4]

TABLE 4. New Dale-Chall score interpretation.

Raw Score	Final Score
$\leq$ 4.9	Grade 4 and Below
5.0 - 5.9	Grade 5 - 6
6.0 - 6.9	Grade 7 - 8
7.0 - 7.9	Grades 9 -10
8.0 - 8.9	Grades 11- 12
9.0 - 9.9	Grades 13 - 15 (College)
> 10	Grades 16 and Above

$$\begin{aligned} RNDC &= 0.1579(DWP) + 0.0496(ASL) \\ DWP &= (\frac{DWC}{WC})100 \\ NDC &= \begin{cases} RNDC + 3.6365, & \text{if } (DWP) > 5 \\ RNDC, & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

ASL is Average Sentence Length, and DWP is Difficult Words Percentage, DWC is Count of Difficult words, and WC is word count.

7) RAYGOR READABILITY FORMULA

Raygor Readability Formula (RRF) is based on the concept of difficult words where words with more than five characters are considered as difficult ones. It was an efficient replacement of Fry based method, where Fry is based on syllable counting, which is more time-consuming. RRF is more suitable for average reading level, targeting US grades 6,7, and 8. The formula is based on 100 word samples, where different 100 words samples can be extracted from longer documents.

$$RRF = 0.284\frac{WC}{SC} + 0.0455\frac{LeC}{WC} - 2.2029 \tag{10}$$

Here LeC represents Letter count, WC and SC annotates word count, and sentence count respectively.

Therefore, on the basis of average number of sentences per 100 words and average number of difficult words, RRF refers to US grade level for which the text would be suitable to read.It is considered to be efficient as well as reliable measure for average readability assessment. However, the results are invalid for elementary as well as advance level texts, and are suitable for the universal readability level.[6]

[6] https://readable.com/blog/the-raygor-readability-graph/

8) TEXT STANDARD

In order to establish the readability consensus based on above measures, PyPI's textstat presented the Text Standard measure. It is a black-box representation of readability consensus[7] python library.

C. FUNCTIONAL LITERACY BASED

UNESCO defines functional literacy as "the ability to identify, understand, interpret, create, communicate and compute"

Unlike already discussed metrics, Forcast is suitable for the non narrative technical content that is the content with incomplete sentences. Since LOs also comprises of Lecture notes and presentations, in which information may be represented as a bulleted list comprising of incomplete sentences, Forcast would also be useful. Forcast formula correlates 0.66 with reading comprehension as measured by reading tests. Forcast score is computed using (11), score between 9 to 10 refers to text with universally accepted readability.

$$ForcastGrade = 20 - \frac{MSC}{10} \tag{11}$$

Here MSC represents MonoSyllable Words Count.

D. ENGLISH L2 READABILITY INDICES

Several formulas have been developed to assess readability levels based on US grade levels. However, in recognition of the needs of L2 (second language) users, only a few metrics have been specifically designed and formulated.

1) LIX

The Swedish research Carl-Hugo Björnsson suggested a stable, less computation intensive, efficient and easy to compute formula for readability. The Lix formula is weighted average of word and sentence factors. Here, word factor maps to the percentage of longer words, that is the words with length greater than 6. The sentence factor is computed using average words per sentences. Björnsson suggested using the 2000 words sample for the desired results. Universal readability is achieved if the Lix measure for the text results in 40 or below.Lix score of 60 refers to very difficult text, whereas text with the Lix score of 20 refers ver easy text.

$$Lix = \frac{WC}{PrC} + \frac{LWC}{WC}100 \tag{12}$$

Here WC represents Word count, and PrC is the period count which is measures using first letter capitalization or full stops, LWC is a count of Longwords in the respective LO.

2) RIX

Syllable counting is an effective mechanism for computing the readability of English text, however, it is not suitable for other languages. Therefore, adopting a measure without syllable counting is an efficient mechanism and can be adopted for other languages as well. Although Lix has better accuracy, objectivity and claims to have better efficiency for assessing readability where English is considered as L2, however, Lix does not specify grade level. A simplified version of the Lix formula, known as Rix score, was introduced to assess readability based on grade level. For universal readability, Rix score of 8 or below is recommended.

$$RIX = \frac{LWC}{SC} \tag{13}$$

Here, LWC specifies count of the words with length > 6 characters and SC represent total number of sentences in LO.

Table [5] summarizes the text difficulty techniques being used in our work.

TABLE 5. Readability indices employed.

Technique	Characteristics and Parameters	CR
AI-LF [43] [23]	Lexical richness(type token ratio),lexical chain span, lexical chain length	-
ARI [32]	Approximate representation of US grade level, efficiency at the cost of reliability (due to syllables)	-
CLRI [44]	Sample text from the passage, efficiency as characters instead of syllables	-
EFLaw	Recommended for L2 Learners, lesser the better,	-
FKEL [41]	One of the frequently used mechanism based on syllables, words, sentences	0.91
FKGR	Developed for general reading, no upper bound adaptation of FKEL	-
Forcast	Developed for non-narrative text	0.35
GFRI	Based on words, sentences length, it estimates the text readability.	0.91
Lix	Difficulty of reading a foreign text (L2 learners) based on longword, words and sentence counts	high
LRRI	weighted average of monosyllables and polysyllables in the text	
NDC [45]	Based on percentage of difficult/ non vocabulary	0.93
Rix	Simplified version of Lix, presents grade level	high
RRF	Based on average of difficult words, average sentences per 100 words	-
SMOG	Based on complex words(polysyllable words) and sentences count, three samples of ten sentences each, take square root, approximate to the nearest perfect square.	0.985

E. LEXICAL MEASURES

In order to analyze text's linguistic richness, lexical diversity (LV) as well as lexical density (LD) are used. 'Lexical diversity' is a measurement of how many unique lexical words exist in a text. Lexical words are words such as nouns, adjectives, verbs, and adverbs that convey meaning in a text. We have assessed the LV of our samples for each datasets. LV is assessed using Type-Token-Ratio. The term token refers to the total number of running words, while the term type refers to the number of distinct word-forms in the text [46]. Root Type Token Ratio (RTTR) takes root forms of the words into consideration instead of simple tokens and then apply TTR. Corrected Type Token Ratio (CTTR) adjusts the TTR value based on the length of the text. Mean segmen-

[7] https://pypi.org/project/textstat/

tal TTR (MSTTR) is computed by splitting the tokens into non-overlapping segments of the given size, TTR for each segment is calculated and the mean of these values is returned. This is helpful to identify lexical diversity of non overlapping text. Moving Average Type Token Ratio (MATTR) is a variation of the Type Token Ratio (TTR) that calculates the average TTR across a sliding window of consecutive words in a text. It provides a measure of lexical diversity that takes into account the variation of vocabulary richness within a text. MATTR is particularly useful in analyzing texts where the lexical richness varies across different sections or where there may be significant differences in vocabulary usage.

Lexical Density (LD) refers to the proportion of content words or lexical items (such as nouns, verbs, adjectives, and adverbs) in a text compared to the total number of words. It measures the degree of lexical information or meaningful content in a given text and is helpful to assess memory retention and thus impacts the readability. The Hypergeometric Distribution Diversity Measure (HDD) is a method that assumes random samples of 42 words. It calculates the probability of encountering any given token for each lexical type in a text. A higher HDD value suggests higher lexical density, indicating a more diverse vocabulary and potentially more complex language use.

The Measure of Textual Lexical Diversity (MTLD) assesses lexical diversity by analyzing the number of words encountered until a specific threshold is reached. It takes into account the number of unique words, word repetitions, and text length. Lower MTLD values indicate higher lexical density, suggesting a more concentrated use of vocabulary.

The Dugast measure of lexical density calculates the ratio of the number of different words (or word types) to the total number of words (or word tokens) in a text. A higher percentage suggests a more diverse vocabulary and greater lexical richness in the text.

Honore's Exponent (HER) is a measure that evaluates the vocabulary richness of a text by examining the frequency of word occurrences. It considers the number of unique words and their frequencies in a text. HER provides a measure of how evenly the vocabulary is distributed, with higher values indicating higher lexical diversity. Lexical Chains were introduced to represent that coherence is a result of cohesion, and is independent of grammatical structure and represent the sequence of related words in the content. [24]

Table [6] represents the measures we have used to assess the lexical richness of lectures, tutorials, course overview in our data sets.

STS stands for successive text segments.

IV. DATASETS

Availability of larger data sets hold a pivotal position in training models using deep learning. Few data sets have been used as benchmark data sets for readability assessments. WeeBit, OneStopEng, CommonLit Ease of Readability Corpus (CLEAR)and Cambridge datasets have been widely used.

TABLE 6. Lexical measures used.

Category	Measure	Measure Description	Formula
Lex-Density	TTR	Type Token Ratio	T/W
Lex-Density	RTTR	Root TTR	$T/\sqrt(W)$
Lex-Density	CTTR	Corrected TTR	$T/\sqrt(2xW)$
Lex-Density	MSTTR	Mean Segmental TTR	Average TTR for STS
Lex-Density	MATTR	Moving Average TTR	Average(TTR(window))
Lex-Diversity	HDD	Hypergeometric Measure	Probability of tokens of lexical type
Lex-Diversity	MTLD	Textual LD	
Lex-Diversity	SLDM	Summer's Measure	log(log(T)) / log(log(W))
Lex-Diversity	DLDM	Dugast's Measure	$(log(w)^2)/(log(W)-log(T))$
Lex-Diversity	HER	Herdan Measure	$: log(terms)/log(words)$
Lex-Chains	MChLen	Length of Largest Chain	
Lex-Chains	TChain	Count of Lex-Chains	

TABLE 7. WeeBit corpus composition.

Grade	Age	Articles	ASA
L2	7 to 8	629	23.41
L3	8 to 9	801	23.28
L4	9 to 10	814	28.12
KS3	11 to 14	644	22.71
GCSE	14 to 16	3500	27.85

A. EXISTING READABILITY BENCHMARK DATASETS

1) WeeBit

WeeBit [12], considered a gold standard in readability analysis, is a combined corpus of WeeklyReader[8] and BBC-Bitsize.[9] WeeklyReader is an educational newspaper targeting children between the ages of 7 and 12 years. It classifies the text, covering diverse topics from science to current affairs, into four categories based on the learner's age. BBC-Bitsize consists of grade-level articles categorized into different Key Stages (KS), including KS1, KS2, KS3, and GCSE.

WeeBit utilizes various features such as lexical, syntactic, and simple count-based measures, including average syllables per word, average sentence length, Flesch Kincaid score, and Coleman-Liau readability indices, to assess readability.

Following table discusses the composition of WeeBit corpus.

ASA refers to average sentences per article.

2) OneStopEng

OneStopEnglish[10] [13] provides a balanced dataset of English language texts categorized into Easy, Average, and Advanced levels. The corpus consists of 189 texts, each available in three versions, resulting in a total of 567 texts. These texts were collected from onestopenglish.com between 2013 and 2016.[11]

OneStopEnglish utilizes a range of features, including classical, lexical, syntactic, psycholinguistic, and discourse-based features such as Coh-metrix and coreference chains from CoreNLP, to train the models. While the corpus incorporates rich features, it is worth noting that the sample

[8] http://www.weeklyreader.com

[9] http://www.bbc.co.uk/bitesize

[10] https://github.com/nishkalavallabhi/OneStopEnglishCorpus

[11] https://onestopenglish.com

TABLE 8. Existing benchmark datasets.

Properties	WeeBit	OneStopEng	Cambridge	CLEAR
Target Audience	General	L2	L2	General
Covered Age	7 - 16	Adult	A2-C2 (CEFR)	7 -16
Class-Balanced?	No	Yes	No	No
Curriculum-Based?	No	No	Yes	Yes
Classes	5	3	5	3
Items per Class	625	189	60	varied
Tokens per Item	217	693	512	varied
Access	Restricted	Public	Public	Public

size is relatively small, and the literature covered in the corpus is not specifically focused on scientific topics.

3) CLEAR

The CLEAR Corpus[12] consists of 4,785 reading passages extracted from English Language Arts classroom contexts for grades 3-12. Each small passage is treated as a separate learning resource within the corpus. The CLEAR Corpus includes various readability indices such as FRE, FK, ARI, SMOG, NDC, CAREC, CAREC_M, and CML2RI.

The corpus serves as a valuable resource for research and development in the field of readability assessment and offers insights into the readability levels of texts used in English Language Arts education.

4) CAMBRIDGE

The Cambridge Corpus[13] is a corpus specifically designed for English language assessment at advanced levels. It comprises a collection of 3,125 articles, which are categorized into five distinct readability levels ranging from age 7 to 16.

The corpus serves as a valuable resource for studying and analyzing the linguistic characteristics and readability of texts targeted at advanced English learners. It provides researchers and educators with a comprehensive dataset to explore language usage and proficiency at different stages of language development.

Following table summarizes the existing readability data sets:

In summary, the discussed corpora, including WeeBit, OneStopEnglish, CLEAR Corpus, and Cambridge Corpus, serve as valuable benchmarks for readability analysis in the English language. They cover a wide range of educational contexts and provide insights into the readability levels of texts used in English Language Arts education. However, it's important to note that these corpora primarily focus on general English language resources and may not specifically cater to technical or domain-specific materials. Researchers and practitioners in the field of Computer Science and readability assessment for CSLR may need to explore or develop specialized corpora to address their specific needs.

B. AGREE DATASET

Although the datasets mentioned above have been extensively used, they are not specifically tailored for research literature. OnestopEnglish, on the other hand, is geared towards adults and divides its text into three categories, but does not encompass scientific literature. To effectively evaluate the level of difficulty of scientific learning resources using readability indices, we have selected seven datasets. These datasets comprise the six existing datasets along with SELRD, our own dataset that represents software engineering resources. The inclusion criteria for these datasets include publicly available, university-level learning materials related to the domains of computer science and/or natural language processing in the English language. The seven datasets we have chosen are Lecturebank (LB), Tutorialbank (TB), University Course Dataset (UCD), NPTEL MOOC dataset, AL-CPL, Software Engineering Learning Resource Dataset (SELRD) and the ACL Anthology Scientific Corpus (AAN).

1) LECTUREBANK

Li [18] presented the manually collected dataset to extract prerequisite relations from the existing learning resources. The dataset comprises 1352 lectures referencing to domains of Natural Language Processing (NLP), Machine Learning(ML), Artificial Intelligence (AI), Deep Learning(DL), and Information Retrieval (IR)from 60 courses referenced from renowned universities.[14] With the vocabulary of 1221 terms, covered in 51939 slides with the total of 2546.65 tokens per lecture. This dataset is rich collection of NLP resources. In addition, the dataset is constantly evolving with LOs being added.

2) TutorialBank

With an objective to facilitate the learning of Natural Language Processing (NLP) by managing the fast changing learning landscapes in the Artificial Intelligence and Deep Learning, Tutorialbank serves as a collection of diversified learning resources of varied pedagogical significance. While Lecturebank comprises of lecture files only, Tutorialbank [19] refers to the manually collected and organized learning resources. These resources include surveys, long papers, tutorials, corpus, blog posts,code bases,libraries, and naclos. With more than 6300 quality controlled resources in first iteration, over 5000 resources in a subsequent contribution,[15] it serves as an updated collection of learning resources.

3) UNIVERSITY COURSE DATASET (UCD)

The rich collection of the courses offered along with their description is aggregated in UCD. It presents the course description of 654 courses. However, it only represents the syllabus to be taught in the courses offered at Princeton, MIT, Stanford, Illinois, Carnegie Melon, Princeton, Maryland, Penn State University(PSU) and Iowa State University.

[12] https://github.com/scrosseye/CLEAR-Corpus/tree/main
[13] https://en.wikipedia.org/wiki/Cambridge_English_Corpus
[14] https://github.com/Yale-LILY/LectureBank
[15] https://github.com/Yale-LILY/TutorialBank

TABLE 9. TutorialBank composition.

Pedagogical Type	Count
Corpus	136
Courses	72
Lectures	126
Libraries	1014
Naclo	154
Tutorial	2044
Surveys	374
Paper	1176
Resource	1065

Average course description of the text is 710 characters.The dataset is incorporated to test the text with lower sentence count. This dataset is useful to assess readability indices for smaller texts.

4) ACL ANTHOLOGY NETWORK CORPUS (AAN)

AAN is a collection of the text of the research articles published by Association of Computational Linguistics (ACL).[16] It is a rich collection of text extracted from 18,290 research articles presented in proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL) since 1979 [20]. With the average content length of 29000 characters per LO, the corpus is inflicted with spelling errors. Being a textual representation, the corpus text does not include formula, diagrams and other visual elements. However, it is useful for assessing the larger texts.

5) NPTEL MOOC

The National Program on Technology Enhanced Learning (NPTEL) dataset comprises of videos transcription on open learning material on science and technology including Computer Science, Biotechnology, Engineering disciplines. It has total of 19500 crawled videos.[17] The sample dataset has over 382 lectures transcripts with vocabulary size of 345, and average transcription size of approximately 28000. [22]

6) AL-CPL

Liang [47] suggested the dataset development using the Wiki Concept Map (WCM) [21]. WCM is the collection of Wiki Concepts based on the concepts acquired from textbooks of Data Mining, Geometry,[18] Physics[19] and Calculus.[20] With 120, 89, 153, and 224 concepts referred in domains of Data Mining, Geometry, Physics, and Pre-Calculus, total of 586 Wikipedia concepts were referred and their data is extracted.

TABLE 10. CS LR datasets.

Dataset	Annot	Domain	LR	Topic Gen	Vocab
LB	Manual	NLP, ML, DL, AI	1352	Auto	1221
TB	Manual	NLP, AI, ML, IR	6300	Manual	17088
ALCPL	Auto	CS	498	Scrapping	21823
AAN	None	CS	18290	-	19167901
NPTEL	Auto	CS	382	List	17997
UCD	Auto	CS	654	List	48635
SELRD	Manual	SE	704	Manual	58342

The summary of the selected datasets is as presented in table [10]:

7) SOFTWARE ENGINEERING DATASET (SELRD)

Most of the dataset pertaining to LOs aggregate and process LR pertaining to subdomain of ML and AI. However, data set for learning Software Engineering(SE) is rarely discussed. In order to fill this gap, we have selected the courses from renowned universities MIT (Open Courseware 16-355j Software Engineering Concepts Offered in fall-2005), (6.170 Software Studio offered in Spring 2013), (1.124J Foundation of Software Engineering offered in Spring 2000 by Prof. Kevin Amaratunga) Rutgers Software Engineering by Ivan Marsic, Virginia Tech (Fall 15 session of CS 3704 by Meng) lectures, book extracts and Wikipedia scrapping of topics found in these courses. Topics extracted books were cataloged section wise from SE books. We have also added the publicly available PowerPoint presentations and books,[21,22,23] on Software Engineering in SELRD.

The AGREE dataset encompasses a wide and diverse collection of Computer Science and Software Engineering Learning Resources (CSLR), consisting of 18,57 topics and 42,850 resources. Various sources were included in the dataset to ensure comprehensive coverage. Lecturebank contributed lecture materials in a bulleted list format, often without full stops, organized under specific headings. AL-CPL provided book excerpts and well-formatted text from Wikipedia. NPTEL contributed transcribed text from video lectures, and AAN offered extracted text from research papers. Tutorialbank contributed a diverse range of learning resources. However, for the purpose of readability assessment, code bases and libraries were excluded. SELRD, the Software Engineering Learning Resources Dataset, includes lectures in PDF and PowerPoint formats, transcripted video lectures, sections from books, and content extracted from Wikipedia, specifically related to Software Engineering (SE) topics.

V. PRE-PROCESSING

Lecturebank [18] as well TutorialBank [19] refers the available lectures, and other learning resources like articles, book

[16] https://aclanthology.org/

[17] https://github.com/AI4Bharat/NPTEL2020-Indian-English-Speech-Dataset

[18] Dan Greenberg, Lori Jordan, Andrew Gloag, Victor Ci-farelli, Jim Sconyers,Bill Zahnerm, "CK-12 Basic Geometry.

[19] Mark Horner, Samuel Halliday, Sarah Blyth, Rory Adams, Spencer Wheaton, "Textbooks for High School Students Studying the Sciences", 2008

[20] Stewart, James, Lothar Redlin, and Saleem Watson. Precalculus: Mathematics for calculus. Cengage Learning,2015.

[21] https: $//iansommerville.com/engineering - software - products/presentations/$

[22] https: $//www.ece.rutgers.edu/ \quad marsic/books/SE/book - SE_marsic.pdf$

[23] https: $//www.craiglarman.com/wiki/index.php?title = Books_byC raig_Larman$

chapters via respective URL. The links referring to libraries, and code snippets cannot to be assessed for standard readability as they are specific computer language dependent and code snippets follow the specific language syntax whereas libraries are in binary formats. We have used these URLs to download the lecture files, surveys and tutorials available in Portable Document Framework (PDF) and PowerPoint(PPTX) file or text transcription of video files. PDF Data was extracted from the respective files using python packages of PyPDF2 and textract whereas python-pptx was used to extract text from PPTX. The content of learning resources referred by these datasets was saved in comma separated values(CSV) format as well as separate text-files. Dataset referred by Wang et al. [21] consists of the topics extracted from text books of different subjects. These topics were used to extract content of respective Wikipedia page using Wikipedia library.[24] White spaces which emerged due to text extraction from existing learning resources (while translating equations or images) were removed. Research papers in AAN were inflicted with too many spaces and new line characters. Those spaces were removed. In addition, comma punctuation in text was replaced by hash sign to make it compatible to csv storage format to ensure reusability, portability and interoperability.

VI. METHODOLOGY

In order to establish the comprehensive readability analysis on scientific literature pertaining to Computer Science and Natural Language Processing, following steps were carried out.

1) **Dataset Selection** Six datasets of Computer Science/ Natural Language domain consisting of LO specified in English Language were selected. Dataset selection criteria includes that language resources must be related to CS. Secondly, LR must be specified in English Language. Moreover, it comprises text or text transcription of multimedia resource
2) **Dataset Generation** Dataset consisting of learning resources of SE was added to extend existing body of knowledge and to facilitate learners of Software Engineering domain.
3) **Dataset Aggregation** Unified format for the datasets was established. Contents from six selected datasets [10] and our designed SELRD were aggregated to yield 42850 LR. In addition, keywords for each LR were extracted using KeyBERT. Common format includes LO ID, Dataset Reference (used for provenance i.e., to trace back from which dataset or book we have extracted this LO), cleaned text, and associated keywords.
4) **Sample Selection** from each of these seven datasets, statistically significant sized samples of around 200 samples, total of 1448 samples were extracted. The sample consists of lecture notes, Wikipedia articles, research papers, and text transcription of video lectures being termed as Learning Objects "LO". The LOs were in text, HTML, PDF and PPTX formats. The sample articles were unified to a common text format.
5) **Gold Standard Annotation** Selected sample of size 1448 was annotated by two annotators for readability assessment. These annotators are graduate students/ PhD Scholars of Computer Science Department.
6) **Readability Assessment using Readability Indices** The sample of size 1448 was assessed using the readability measures mentioned [5]. The US grade score was computed from readability measures, and was standardized into three classes: Easy, Average, Difficult.
7) **Readability Assessment using Lexical Analysis** The text was also assessed using the lexical measures. It includes evaluation of text to check Lexical Density, Lexical Diversity and Count and span of Lexical Chains mentioned in Table [6].
8) **Ensemble Results** The readability index measures were ensemble using mean, median and mode. The median and mode has higher correlation with gold standard as compared to mean.
9) **Analyze Results** The results of ensemble models were analyzed using correlations, standard deviation, and accuracy measures.
10) **Applying ML Models** The dataset was split into training and testing dataset using random state= 42 to get the same train and test sets across different executions. Later, The readability dataset was trained using the machine learning (lazy classifier[25]) models. Metrics for accuracy, F1 score and efficiency in terms of time taken to compute the results were recorded.

VII. GOLD STANDARD ANNOTATION

A sample set comprising of 1448 samples (around 200 samples from each dataset) was annotated by two annotators, PhD scholars of Computer Science domain. The annotators manually assigned a readability score to each item in the set, classifying them as "easy", "average", or "difficult", which was recorded as the gold standard. To reduce subjectivity, the gold standard was established based on factors such as the content's length, existing knowledge of the topic, acronyms used, proper nouns, and acronyms. The coupling factor was also taken into account, which indicates whether a learning object (LO) has a dependency on another LO, where i and j are subscripts representing different LOs. This occurs if the LO refers directly to other LOs or is a continuation of other LOs. Inter-annotator agreement was established using the Cohen Kappa measure, resulting in a score of 0.87 showing near perfect agreement.[26] Table 11 specifies the features considered for the Gold Standard.

[24] https://pypi.org/project/wikipedia/

[25] https://pypi.org/project/lazypredict/

[26] https://towardsdatascience.com/interpretation-of-kappa-values-2acd1ca7b18f

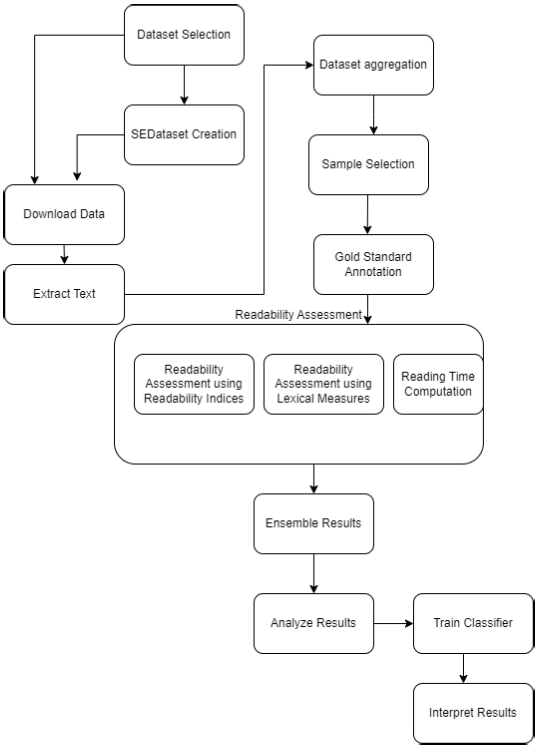

FIGURE 1. AGREE dataset design methodology.

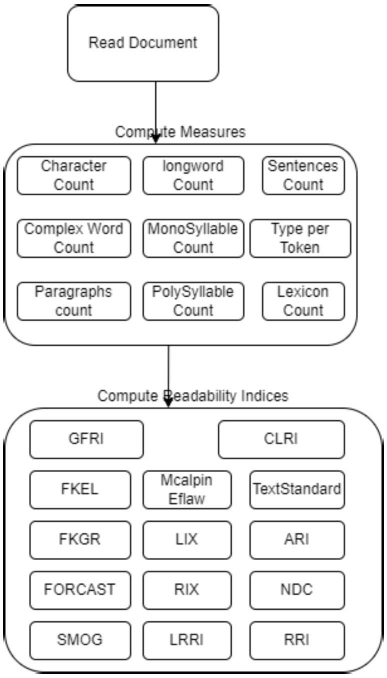

FIGURE 2. Readability Indices Computation describes the existing readability Indices applied an being used in ensemble model.

A. BENCHMARK CRITERIA

In the literature, certain characteristics have been suggested for a benchmark dataset [48].We evaluated AGREE with respect to these characteristics.

TABLE 11. Gold standard evaluation parameters.

Feature	Annotation Criteria
Acronyms	How extensively acronyms or abbreviations are used
Content's Length	Word count of the Learning Resource (LR)
Coupling Factor	How Frequently the sections or tables are co-referenced
Clarity	Whether the text is written clearly.
Proper Nouns (PN)	Higher the PN count, more difficult it is
Knowledge of the topic	Existing topic knowledge makes content easier to understand.
Introduction	Well presented introduction for each section
Summarizing	Well written summary of each section

1) Relevance
The AGREE dataset includes a wide range of learning resources such as books, lectures from renowned universities, research papers, and Wikipedia content related to CS topics discussed in lectures. These LOs are relevant for individuals looking to learn about computer science and can be used in recommender systems. Furthermore, the dataset includes information about the difficulty level and reading time for each LO.

2) Representativeness
refers to the extent to which a dataset covers the full range of possible events it is intended to represent. The AGREE dataset includes a wide range of resources with diverse pedagogical significance, including book sections, Wikipedia articles, research papers, and publicly available online lectures in PDF/PPTx formats. Therefore, the dataset is considered to be highly representative of the domain of computer science.

3) Non-Redundancy
To reduce redundancy from our dataset, we ensured that learning resources (LRs) are included only once in a particular format. However, there exist scenarios where the same section of the book is covered in PowerPoint slide. However, since both have different level of description and format, we have incorporated both the versions as LOs.

4) Scalability
In the future, we plan to expand the dataset by including data that specifically represents definitions, examples, and case studies related to the existing topics. This ensures that the dataset can scale both vertically and horizontally, accommodating the growth of data in terms of both volume and diversity.

5) Reusability
The dataset provides metadata related to the content and readability measures in comma-separated values (CSV) file format. Being an open dataset, it can be accessed by the public and reused in various scenarios, such as curriculum design and machine learning

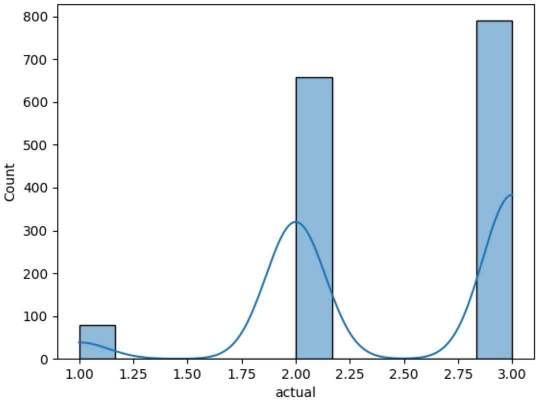

FIGURE 3. Data distribution among difficulty classes in the sample set.

algorithms for recommender systems. Its accessibility and openness make it valuable for multiple applications.

We have evaluated our dataset with respect to benchmark characteristics. Although we have aggregated the LOs, and we have ensured that duplication of resources is prevented. In addition, CSV file format, being portable and interoperable, is used to store the results.

VIII. RESULTS AND DISCUSSION

Initially on the statistically significant sample size, we conducted the exploratory data analysis. The data shows that sample is multimodal distribution, with more literature referring to the average and difficult level and few LOs referring to the easy level. This is in accordance with our objective of selecting the learning objects.

We carried out the following experiments, using Jupyter 6.5.2 on Intel Core i7 machine with MS Windows 11 to seek answers to our research questions.

RQ1: In order to assess the consensus between the readability formula, we first computed the readability indices presented in Table [5] for samples selected from AGREE [10]. Each computed value was interpreted according to the respective scale of readability index, and was assigned the US grade level. Later, we computed Standard deviation (SD) using results of readability Indices for each LO in the sample. Table [12]summarizes the SD values of LO of samples selected. Results show that standard deviation of LOs referring to transcripted video lectures (NPTEL) and research papers (AAN), (having LOs with large token count per LO) is very small and ranges between 0.302 to 0.905. The smaller SD values shows that computed readability indices are clustered towards its mean and therefore shows consensus.

RQ2:

In order to assess the consensus of computed readability indices (RI) with the gold standard, we first computed RI for each selected LO [5]. Results of each RI were interpreted and difficulty level was assessed using universal readability level (RL), and the values above and below RL being categorized as difficult and easy LO. Later, the computed difficulty level values of each RI of each LO, were ensembled using Mean, Mode and Median. Table 13 shows that ensemble values of readability indices and manually annotated gold standard's readability values have consensus.

TABLE 12. SD values to assess consensus among readability indices.

Dataset	Min- SD	Max- SD
AAN	0.302	0.905
AL-CPL	0.54	1.01
LB	0.52	0.89
NPTEL	0	.809
TB	0.48	0.96
UCD	0.603	1.1

TABLE 13. Accuracy of readability indices vs gold standard.

Dataset	Mean Vs Gold	Median Vs Gold	Mode Vs Gold
AAN	0.63	0.83	0.70
AL-CPL	0.80	0.76	0.8
LB	0.90	0.80	0.83
NPTEL	0.88	0.92	0.92
TB	0.73	0.77	0.77
UCD	0.80	0.77	0.76

TABLE 14. Correlation of reading time (RT) with readability measures.

Readability Measure	Correlation Co-efficient
Lexicon Count Vs RT	0.99
Total Lexical Chains Vs Rt	0.93
Max. Lexical Chain Length Vs RT	0.66
RTTR Vs RT	0.54

RQ3: In order to assess the relationship between reading time and readability measures, we first computed reading time using average reading of 14.69ms per character for non-fiction text. [49] Later, the correlation coefficients were computed to assess the strength of the relationship between reading time and measures of readability Indices. In addition, reading time's correlation with lexical measures was also evaluated. Table [14] shows the correlation values between readability measures and the reading time (RT)

Table 14, Fig [4] show that Lexicon count, Lexical Chain count. Chain length has higher positive correlation with RT, showing that when lexicon count or lexical chains count or maximum length of the lexical chain increases, reading time will also increase. In addition, the higher correlation between Root Type Token Ratio (RTTR) with RT shows that when unique tokens (types) increases (when LO comprises non-repeating and unique words) reading time also increases.

Therefore, by computing lexicon count, count of chain lengths, and Root Type Token ration (RTTR) reading time can also be estimated.

The positive correlation in Fig [4] shows the direct relationship between reading time and Lexicon count, count of lexical chains, root type token ratios (RTTR) and maximum

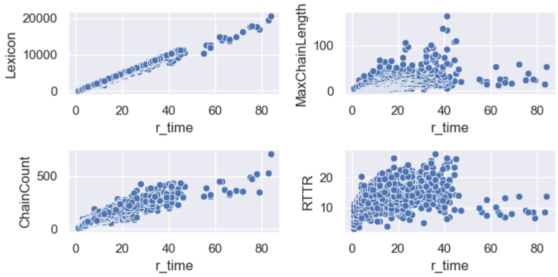

FIGURE 4. Correlation between reading time and readability measures.

TABLE 15. Classifiers results on AGREE.

Model	Accuracy	F1 Score	Time Taken
LGBMClassifier	0.83	0.83	0.5
RandomForestClassifier	0.82	0.82	0.42
AdaBoostClassifier	0.78	0.78	0.24
BaggingClassifier	0.8	0.8	0.19
DecisionTreeClassifier	0.79	0.79	0.03
ExtraTreesClassifier	0.82	0.81	0.22
LinearDiscriminantAnalysis	0.78	0.78	0.36
LogisticRegression	0.77	0.77	0.06
ExtraTreeClassifier	0.73	0.73	0.02
QuadraticDiscriminantAnalysis	0.46	0.54	0.09
SVC	0.81	0.8	0.1
Perceptron	0.81	0.79	0.02
SGDClassifier	0.76	0.76	0.04
LabelPropagation	0.73	0.73	0.14
LabelSpreading	0.73	0.73	0.14
CalibratedClassifierCV	0.78	0.77	1.38
KNeighborsClassifier	0.75	0.74	0.06
LinearSVC	0.78	0.77	0.35
PassiveAggressiveClassifier	0.77	0.76	0.02
RidgeClassifierCV	0.79	0.77	0.05
BernoulliNB	0.55	0.59	0.03
RidgeClassifier	0.78	0.76	0.02
NearestCentroid	0.58	0.62	0
GaussianNB	0.22	0.27	0
DummyClassifier	0.45	0.28	0.02

chain length with few outliers due to variation in length of text in the corpus.

RQ4:

The problem of assessing the readability of the text is inherently a regression problem where we need to predict the difficulty level of a document as [0,1] with 0 being the easy one and 1 representing difficult one.

But the ground truth for regression is hard to gather, atleast for now, whereas we can make k bins of the regression labels, it will become k-class classification problem with k=3 representing 3 classes of easy, difficult and average readability level of text. Formally,

$$y_i \in \{0, 1, 2\} \tag{14}$$

Now given a set, D, of N documents, $D = \{D_1, D_2, \ldots, D_n\}$ consider $\mathbf{x_i}$ is a feature vector representing the i_{th} document as a point in some d-dimensional feature space. The feature vectors could be contextual embedding computed through BERT or non-contextual ones like TFIDFs. The classification problem can then be modeled as a parametric function $f_{\mathbf{w}}$, with parameters $\mathbf{w}$, that maps each feature vector to it's corresponding class label. More formally, if $f_{\mathbf{w}}(\mathbf{x_i})$, is the class prediction made by the model against the document D_i, represented by the feature vector $\mathbf{x_i}$, then we need the parameters $\mathbf{w^x}$, that minimize a loss function L over all training examples as follows:

$$\mathbf{w^x} = arg \min_{\mathbf{w}} \left\{ \sum_{i=1}^{n} L(f_{\mathbf{w}}(\mathbf{x_i}), y_i) \right\}$$

depending upon choice of the parametric model $f_{\mathbf{w}}$ and loss function L, different classification algorithms can behave differently on the problem at hand.

In order to assess the suitability of classifiers for our dataset, we used “scikit learn”’s “lazy predict”. The results for AGREE dataset in current configuration shows the at out of 25 classification algorithms, the Light Gradient Boost Method (LGBMClassifier), Random Forests (RF), Extra Trees Classifier are most suitable for readability classification for our dataset with accuracy 0.83 and 0.82, 0.82 respectively. LGBM Classifier is efficient and supports parallelism for larger datasets. Random Forest is also scalable and can handle large datasets with higher accuracy. Extra trees classifier aggregates the multiple de-correlated decision trees results “forest” to yield the class label. However, Extra tree classifier has relatively higher accuracy with F1 score and is more efficient. Therefore, we recommend classifying readability level of AGREE using ExtraTreesClassifier.

RQ5: The existing benchmark datasets, such as WeeBit, OnestopEnglish, and CLEAR, are commonly used to evaluate the readability of English text for general purposes. However, these datasets may not be sufficient for assessing the readability of CS text comprising lectures, research papers and other technical material. This highlights the need for a dataset specifically designed for assessing the readability of computer science-related scientific texts.

5 summarizes the tests conducted to answer the research questions.

Our work stands out from existing benchmark datasets like CLEAR, WeeBit by introducing a novel dataset specifically designed for the learning resources of Computer Science (CSLR). This dataset comprises more than 42,000 CSLR, alongwith 26 readability measures, providing a substantial and comprehensive collection of materials relevant to the domain of Computer Science.

One notable distinction of our dataset is the unique structure of the lectures. Unlike research papers, books, or Wiki articles, which are often carefully paraphrased and structured with complete sentences, lectures tend to present information in incomplete sentences, reflecting the dynamic and conversational nature of educational presentations.

By capturing this distinct characteristic of lectures, our dataset offers a valuable resource for studying and evaluating readability in the context of CS education. It provides researchers and practitioners with a more realistic representation of the language and content commonly encountered in

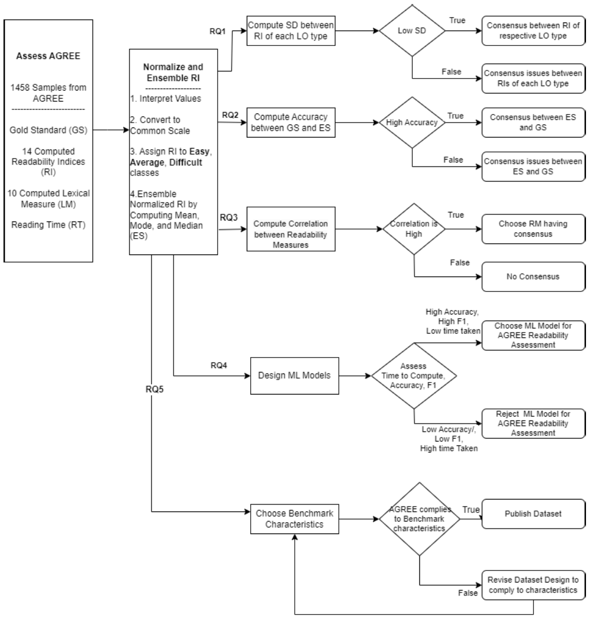
Assess AGREE
1458 Samples from AGREE
Gold Standard (GS)
14 Computed Readability Indices (RI)
10 Computed Lexical Measure (LM)
Reading Time (RT)
Normalize and Ensemble RI
1. Interpret Values
2. Convert to Common Scale
3. Assign RI to Easy, Average, Difficult classes
4.Ensemble Normalized RI by Computing Mean, Mode, and Median (ES)
RQ1
Compute SD between RI of each LO type
Low SD
True
Consensus between RI of respective LO type
False
Consensus issues between RIs of each LO type
RQ2
Compute Accuracy between GS and ES
High Accuracy
True
Consensus between ES and GS
False
Consensus issues between ES and GS
RQ3
Compute Correlation between Readability Measures
Correlation is High
True
Choose RM having consensus
False
No Consensus
RQ4
Design ML Models
Assess Time to Compute, Accuracy, F1
High Accuracy, High F1, Low time taken
Choose ML Model for AGREE Readability Assessment
Low Accuracy/, Low F1, High time Taken
Reject ML Model for AGREE Readability Assessment
RQ5
Choose Benchmark Characteristics
AGREE complies to Benchmark characteristics
True
Publish Dataset
False
Revise Dataset Design to comply to characteristics

FIGURE 5. Summary of the experiments conducted to answer research questions.

CS learning materials, enabling the development and refinement of readability assessment techniques tailored to this specific domain. The AGREE framework holds significant potential in the realm of state-of-the-art transformer-based approaches that heavily rely on data for enhanced readability assessment. By incorporating AGREE into these advanced models, we can achieve improved accuracy in evaluating the readability of various textual resources.

The AGREE dataset uses diverse readability indices and lexical measures to assess text readability. Furthermore, our trained classifier can be used to label new learning objects by predicting their readability level. The dataset was also evaluated based on the characteristics of benchmark datasets discussed in the literature.

Experiments reveal that accuracy measures yield significant results for the larger texts extracted from video lectures of NPTEL, research papers from AAN. For smaller texts (UCD dataset), when TTR is high due to high count of unique words, readability results yield inaccuracy. In addition, results show that counts of lexicon, proper nouns, lexical chain and RTTR positively correlates with reading time. In terms of efficiency, the character and sentence count based readability indices such as ARI, CLRI were efficient to compute as compared to the syllable counts and lexical indices.

Flesch Indices can be easily computed using commonly used word processors like MS Word. "Text standard" readability measure, which was designed to incorporate standardization of readability metrics, is provided as a blackbox measure, and we could not improve the consensus of readability indices using Text Standard measure of textstat. Although Forcast has been designed for non-narrative writing, but it did not provide strong correlation with the lectures presented in PPTX format. The New Dale-Chall list comprises the vocabulary of US grade 4 student, which needs enhancement as some of the very common words like computer, virus, internet and other prevalent words are missing. In order to improve Dale-Chall readability index accuracy, an updated list is recommended. The updated list may also be customized for L2 learners or may be domain specific lists can be established. We have also observed that the role of proper nouns and acronyms in assessing readability may also be incorporated in readability measurement. In addition, we did not opt for Cohmetrix[27] as it imposes the restriction of 15000 characters, does not accept irregular characters. In addition, the errors are also not much intuitive in nature.

Our learning objects comprise the text extracted from digital contents which do not cover the mathematical representations appropriately and therefore accuracy of the readability results were approximated. Mechanisms considering count and complexity of equation, charts, and figures may be incorporated in future to improve readability assessment.

We have also trained the classifiers to generate the model for classifying the readability class of learning objects. Results show that ensembled models Extra Tree classifier has better accuracy, F1 score and efficiency.

In summary, the creation of a computer science-related dataset for readability assessment not only provides targeted resources to learners in this specialized field, but also contributes to the ongoing enhancement of learning materials. By catering to the diverse needs of learners and enabling continuous evaluation and improvement, the dataset plays a pivotal role in fostering effective and high-quality learning experiences in the domain of computer science not only by guiding the learners about the difficulty level of the text but also by providing dataset for the researches in recommender system, text simplification and reading list generation.

IX. FUTURE WORK AND CONCLUSION

In this study we have contributed the AGREE data set comprising learning resources of Computer Science domain which not only serve as learning repository but is also useful for recommender systems, curriculum design, and training deep learning models.It can also be used for learning analytics where predictive modeling can be used to gauge the students' performance in advance. We have also contributed SELRD to cover the domain of Software Engineering, in addition to the AGREE dataset comprising CSLR with their RI, LM and reading time. We have studied the readability of text extracted from these LR as a factor of content, and lexical measures in scientific and research. We have also contributed that for the given configuration of AGREE, extra tree classifier accurately and efficiently models the readability class of the learning objects.

In the future, we plan to scale up the size and diversity of the AGREE dataset by adding more examples and case studies related to the CS domain. In addition, we also plan

[27] http://tool.cohmetrix.com/

to incorporate multi-modality by adding images and audios. We also aim to expand the dataset by including measures related to content formatting, composition, and legibility features such as tables, charts, equations and images. Additionally, we would assess the correlation between readability and text legibility features.

ACKNOWLEDGMENT

The authors would like to acknowledge Dr. Shahzad Sarwar (Professor & Chair, Department of Computer Science, University of the Punjab) for his persistent valuable support. In addition, the support provided by Dr. Muhammnad Kamran Malik (Professor, Department of Data Science, University of the Punjab), Kashif Murtaza (Assistant Professor, Department of Information Technology), Esha Aftab (Assistant Professor, Department of Computer Science) and Dr. Waheed Iqbal (Associate Professor, Department of Data Science, University of Punjab) is highly acknowledged.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] C. Gurajena, E. Mbunge, and S. G. Fashoto, "Teaching and learning in the new normal: Opportunities and challenges of distance learning amid COVID-19 pandemic," *Int. J. Educ. Teach.*, vol. 1, no. 2, pp. 9–15, 2021.

[2] S. L. Schneider and M. L. Council, "Distance learning in the era of COVID-19," *Arch. Dermatological Res.*, vol. 313, no. 5, pp. 389–390, Jul. 2021.

[3] A. A. Funa and F. T. Talaue, "Constructivist learning amid the COVID-19 pandemic: Investigating students' perceptions of biology self-learning modules," *Int. J. Learn., Teach. Educ. Res.*, vol. 20, no. 3, pp. 250–264, Mar. 2021.

[4] P. Plavén-Sigray, G. J. Matheson, B. C. Schiffler, and W. H. Thompson, "The readability of scientific texts is decreasing over time," *eLife*, vol. 6, Sep. 2017, Art. no. e27725.

[5] D. P. Hayes, "The growing inaccessibility of science," *Nature*, vol. 356, no. 6372, pp. 739–740, Apr. 1992.

[6] K. Ito, "Development and update of ATOS," *JR East Tech. Rev.*, no. 19, pp. 52–55, 2011.

[7] J. R. Bormuth, "Readability: A new approach," *Reading Res. Quart.*, vo. 1, pp. 79–132, Apr. 1966.

[8] G. R. Klare, "Assessing readability," *Reading Res. Quart.*, vol. 10, no. 1, pp. 62–102, 1974.

[9] H. Aliakbarpour, M. T. Manzuri, and A. M. Rahmani, "Improving the readability and saliency of abstractive text summarization using combination of deep neural networks equipped with auxiliary attention mechanism," *J. Supercomput.*, vol. 78, pp. 1–28, Feb. 2022.

[10] W. Li, Z. Wang, and Y. Wu, "A unified neural network model for readability assessment with feature projection and length-balanced loss," 2022, *arXiv:2210.10305*.

[11] B. W. Lee, Y. S. Jang, and J. Hyung-Jong Lee, "Pushing on text readability assessment: A transformer meets handcrafted linguistic features," 2021, *arXiv:2109.12258*.

[12] S. Vajjala and D. Meurers, "On improving the accuracy of readability classification using insights from second language acquisition," in *Proc. 7th Workshop Building Educ. Appl. Using NLP*, 2012, pp. 163–173.

[13] S. Vajjala and I. Lučić, "OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification," in *Proc. 13th Workshop Innov. Use NLP Building Educ. Appl.*, 2018, pp. 297–304.

[14] H. Yannakoudakis, T. Briscoe, and B. Medlock, "A new dataset and method for automatically grading ESOL texts," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 180–189.

[15] S. Crossley, A. Heintz, J. S. Choi, J. Batchelor, M. Karimi, and A. Malatinszky, "A large-scaled corpus for assessing text readability," *Behav. Res. Methods*, vol. 55, no. 2, pp. 1–17, 2022.

[16] S. Ghosh, S. Sengupta, S. K. Naskar, and S. K. Singh, "FinRAD: Financial readability assessment dataset-13,000+ definitions of financial terms for measuring readability," in *Proc. 4th Financial Narrative Process. Workshop*, 2022, pp. 1–9.

[17] R. P. L. Buse and W. R. Weimer, "Learning a metric for code readability," *IEEE Trans. Softw. Eng.*, vol. 36, no. 4, pp. 546–558, Jul. 2010.

[18] I. Li, A. R. Fabbri, R. R. Tung, and D. R. Radev, "What should I learn first: Introducing lecturebank for NLP education and prerequisite chain learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 6674–6681.

[19] A. R. Fabbri, I. Li, P. Trairatvorakul, Y. He, W. T. Ting, R. Tung, C. Westerfield, and D. R. Radev, "TutorialBank: A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation," 2018, *arXiv:1805.04617*.

[20] D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara, "The ACL anthology network corpus," *Lang. Resour. Eval.*, vol. 47, no. 4, pp. 919–944, Dec. 2013.

[21] S. Wang, A. Ororbia, Z. Wu, K. Williams, C. Liang, B. Pursel, and C. L. Giles, "Using prerequisites to extract concept maps from textbooks," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2016, pp. 317–326.

[22] M. S. Ananth, "National programme on technology enhanced learning (NPTEL): The vision and the mission," in *Proc. IEEE Int. Conf. Technol. Educ.*, Jul. 2011, p. 8.

[23] L. Feng, N. Elhadad, and M. Huenerfauth, "Cognitively motivated features for readability assessment," in *Proc. 12th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, 2009, pp. 229–237.

[24] J. Morris and G. Hirst, "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," *Comput. Linguistics*, vol. 17, no. 1, pp. 21–48, Mar. 1991.

[25] H.-C. Tseng, H.-C. Chen, K.-E. Chang, Y.-T. Sung, and B. Chen, "An innovative bert-based readability model," in *Proc. Innov. Technol. Learn., 2nd Int. Conf.* Cham, Switzerland: Springer, Dec. 2019, pp. 301–308.

[26] J. M. Imperial, "Bert embeddings for automatic readability assessment," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, 2021, pp. 611–618.

[27] J. Yu, Y. Wang, Q. Zhong, G. Luo, Y. Mao, K. Sun, W. Feng, W. Xu, S. Cao, K. Zeng, Z. Yao, L. Hou, Y. Lin, P. Li, J. Zhou, B. Xu, J. Li, J. Tang, and M. Sun, "MOOCCubeX: A large knowledge-centered repository for adaptive learning in MOOCs," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 4643–4652.

[28] K. Kyle, S. Crossley, and C. Berger, "The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0," *Behav. Res. Methods*, vol. 50, no. 3, pp. 1030–1046, Jun. 2018.

[29] K. Kyle, "Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication," Ph.D. dissertation, Georgia State Univ., Atlanta, GA, USA, 2016, doi: 10.57709/8501051.

[30] S. Zhou, H. Jeong, and P. A. Green, "How consistent are the best-known readability equations in estimating the readability of design standards?" *IEEE Trans. Prof. Commun.*, vol. 60, no. 1, pp. 97–111, Mar. 2017.

[31] J. P. Kincaid, R. P. Fishburne Jr., R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Nav. Tech. Training Command Millington TN Res. Branch, Millington, TN, USA, Tech. Rep. AD A006655, Feb. 1975.

[32] R. J. Senter and E. A. Smith, "Automated readability index," Aerosp. Med. Res. Lab., Univ. Cincinnati, Oh, USA, Tech. Rep. AD 667273, 1967.

[33] W. S. Gray and B. E. Leary, *What Makes a Book Readable*. Chicago, IL, USA: Univ. of Chicago Press, 1935.

[34] A. Boudjella, M. Sharma, and D. Sharma, "Non-native English speaker readability metric: Reading speed and comprehension," *J. Appl. Math. Phys.*, vol. 5, no. 6, pp. 1257–1268, 2017.

[35] C. Brantmeier, "Statistical procedures for research on l2 reading comprehension: An examination of anova and regression models," *Reading Foreign Lang.*, vol. 16, no. 2, pp. 51–69, 2004.

[36] L. Feng, *Automatic Readability Assessment*. New York, NY, USA: City Univ. of New York, 2010.

[37] A. Kiselnikov, D. Vakhitova, and T. Kazymova, "Coh-metrix readability formulas for an academic text analysis," in *Proc. IOP Conf. Mater. Sci. Eng.*, vol. 890, no. 1. Bristol, U.K.: IOP Publishing, 2020, Art. no. 012207.

[38] M. Xia, E. Kochmar, and T. Briscoe, "Text readability assessment for second language learners," 2019, *arXiv:1906.07580*.

[39] C.-H. Björnsson, "Readability of newspapers in 11 languages," *Reading Res. Quart.*, vol. 18, pp. 480–497, Jul. 1983.

[40] K. M. Sheehan, M. Flor, D. Napolitano, and C. Ramineni, "Using *TextEvaluator* to quantify sources of linguistic complexity in textbooks targeted at first-grade readers over the past half century," *ETS Res. Rep. Ser.*, vol. 2015, no. 2, pp. 1–17, Dec. 2015.

[41] W. H. DuBay, *Smart Language: Readers, Readability, and the Grading of Text*. Brussels, Belgium: ERIC, 2007.

[42] R. Gunning, *The Technique of Clear Writing*. New York, NY, USA: McGraw-Hill, 1952.

[43] E. Pitler and A. Nenkova, "Revisiting readability: A unified framework for predicting text quality," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2008, pp. 186–195.

[44] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring," *J. Appl. Psychol.*, vol. 60, no. 2, pp. 283–284, Apr. 1975.

[45] E. Dale and J. S. Chall, "A formula for predicting readability: Instructions," *Educ. Res. Bull.*, vol. 18, pp. 37–54, Jul. 1948.

[46] A. Mauranen and P. Kujamäki, *Translation Universals: Do They Exist?* vol. 48. Amsterdam, The Netherlands: John Benjamins Publishing, 2004.

[47] C. Liang, J. Ye, S. Wang, B. Pursel, and C. L. Giles, "Investigating active learning for concept prerequisite learning," in *Proc. EAAI*, 2018, pp. 1–7.

[48] A. Sarkar, Y. Yang, and M. Vihinen, "Variation benchmark datasets: Update, criteria, quality and applications," *Database*, vol. 2020, pp. 1–16, Jan. 2020.

[49] V. Demberg and F. Keller, "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity," *Cognition*, vol. 109, no. 2, pp. 193–210, Nov. 2008.

MUDDASSIRA ARSHAD (Graduate Student Member, IEEE) received the M.Sc. degree in computer science from Quaid-i-Azam University, Islamabad, Pakistan, in 2001, and the M.Phil. degree in computer science from the University of the Punjab, Pakistan, in 2015, where she is currently pursuing the Ph.D. degree with the Department of Computer Science. She is also serving as an Assistant Professor with the Department of Software Engineering, University of the Punjab. She was a Lecturer with Quaid-i-Azam University for a decade. Her current research interests include natural language processing and human–computer interaction.

MUHAMMAD MURTAZA YOUSAF received the Ph.D. degree from the University of Innsbruck, Austria, in 2008. He worked on networks for grid computing during his Ph.D. study. He is currently a Professor with the Department of Software Engineering, University of the Punjab, Lahore, Pakistan. His current research interests include the transport layer of networks, parallel and distributed computing, cloud computing, data science, and interdisciplinary research.

SYED MANSOOR SARWAR (Member, IEEE) received the Ph.D. degree in computer engineering from Iowa State University, Ames, IA, USA. He started his academic career in April 1982 as a Lecturer at the Department of Electrical Engineering, University of Engineering and Technology (UET), Lahore. From 1988 to 1990, he served as an Assistant Professor at the Department of Electrical and Computer Engineering, Kuwait University. In 1991, he was the Area Chair of computer science at the Pak–American Institute of Management Sciences, Lahore. He served the Lahore University of Management Sciences (LUMS) as a Professor and the Head of the Department of Computer Science. He served the University of Portland as an Assistant Professor and an Associate Professor for over a decade. He served the Punjab University College of Information Technology (PUCIT) for over 12 years and took retirement after serving for more than a decade. He joined UET in 2018.