Published by The College of Science & Information Technology (COSIT), TASFUED, Vol. 19, No. 2, pp. 47-62.

COSIT, TASFUED
Journal of Science and Information Technology (JOSIT)

Incorporating Conceptual Clarity and Topic Coverage into Flesch-Kincaid Readability Model for University Course Content Analysis

[1]Osofuye, O.; [1]Onwusiribe, C.; [1]Sogunle, O.; [2]Osofuye A.; [1]Ogunyale O.; [1]Jegede T.; [1]Abiodun, T.

[1]Department of Computer and Information Sciences, Covenant University, Nigeria.
[2]Business Risk and Internal Control, CapitalSage Holdings, Nigeria.

Corresponding Author: odunayo.osofuye@covenantuniversity.edu.ng
Other Authors: onwusiribe.chioma@stu.cu.edu.ng; olubusolami.sogunle@stu.cu.edu.ng; a.osofuye@capitalsage.ng; ogunyale.oluwaseun@covenantuniversity.edu.ng; tope.jeged@covenantuniversity.edu.ng; theresa.abiodun@covenantuniversity.edu.ng

Abstract

While the Flesch-Kincaid (FK) formula remains a standard for readability assessment, current research highlights its limitations in measuring actual text comprehension due to its focus on surface-level linguistic features. This study addresses this gap by developing the Conceptual Clarity and Topic Coverage (CCTC) enhanced Flesch Reading Ease (FRE) model, which integrates semantic coherence and syllabus alignment into the traditional readability framework. The system, implemented via a Flask-based web application, automates the evaluation of university course materials using a pipeline of Latent Dirichlet Allocation (LDA) for topic distribution and TF-IDF cosine similarity for conceptual clarity. Validation was performed using a dataset of Computer Science courseware benchmarked against expert human ratings. Statistical analysis via ANOVA and Pearson correlation demonstrated that the CCTC-enhanced FRE score achieves a significantly higher correlation with human judgment (τ=0.86, ρ<0.05) compared to traditional metrics such as the standard FRE (τ=0.86), Gunning Fog (-0.56), and SMOG (-0.77). These findings provide empirical evidence that incorporating semantic and thematic depth significantly improves the accuracy of automated readability assessments in technical academic contexts. This research offers a robust framework for educational quality assurance and data-driven curriculum design.

Keywords: CCTC-enhanced FRE Score, Educational Quality Assessment, Readability Metrics, Course Content Analysis, Natural Language Processing

INTRODUCTION

Readability is defined as a linguistic metric that quantifies the ease with which a reader can process and understand a written document (Meng et al., 2020). Within the field of pedagogy, content analysis is frequently utilized to scrutinize instructional materials and student outputs to ensure educational alignment (Teachers Institute, 2024). This method allows researchers to organize vast amounts of

Cite as:

Osofuye, O.; Onwusiribe, C.; Sogunle, O.; Osofuye A.; Ogunyale O.; Jegede T.; Abiodun, T. (2025). Incorporating Conceptual Clarity and Topic Coverage into Flesch-Kincaid Readability Model for University Course Content Analysis *Journal of Science and Information Technology (JOSIT)*, Vol. 19 No. 2, pp. 47-62.

information into thematic categories to identify recurring patterns (Rucks-Ahidiana, 2024). In higher education settings, the effectiveness of these materials is largely dependent on how well complex ideas are communicated to the learner (Hodges et al., 2020). However, accurately measuring this clarity remains a significant challenge due to the limitations of traditional linguistic tools (Choi & Crossley, 2022). Furthermore, the mastery of academic vocabulary is crucial for undergraduate success, yet students often struggle with the transition to specialized academic language (Chung et al., 2024).

Standard assessment formulas, such as the Flesch-Kincaid Grade Level, primarily focus on surface-level features like sentence length and syllable counts (Crossley et al., 2021). Recent studies highlight that these metrics lack a strong correlation with actual cognitive reading performance (Tanprasert & Kauchak, 2021). Furthermore, traditional indices can be easily manipulated to improve scores without actually enhancing the clarity of the underlying content (Vajjala, 2021). This creates a reliability gap when evaluating technical texts, as these formulas often penalize essential domain-specific vocabulary as being too difficult (Ehara, 2023). Consequently, a text might be rated as inaccessible simply because it uses the correct terminology required by the subject matter (Ha & Yaneva, 2018). Linguistic research confirms that the mere presence of long words does not always equate to high difficulty if the reader has specific domain exposure (Bulté & Housen, 2012).

Modern research has begun to pivot toward Natural Language Processing (NLP) to capture document-level semantics and structural dependencies better (Priebe et al., 2012). While deep learning models provide high accuracy in general contexts, they often struggle with the data scarcity found in specialized academic domains (Cai et al., 2020). Other approaches use Latent Dirichlet Allocation (LDA) to provide a more modular and interpretable "white-box" view of text topics (Blei et al., 2003). To improve these models, researchers have investigated normalized approaches for finding optimal topic numbers to ensure semantic stability (Hasan et al., 2021). Also, advanced term-weighting strategies, such as TF-IDF, are now being modified to be more sensitive to word relevance in specific document types (Jalilifard et al., 2020). There is a clear need for a framework that validates whether the complexity of a word is justified by its presence in the syllabus (Alexander et al., 2022).

This study introduces the CCTC-enhanced FRE model, which integrates Latent Dirichlet Allocation (LDA) and TF-IDF to align traditional readability scores with syllabus relevance. By identifying anchor words essential to the curriculum, the model effectively reduces the academic penalty for technical vocabulary. Validated against 114 Computer Science modules, our framework demonstrates a superior correlation with expert judgment (τ=0.86) compared to standard metrics. This research provides a functional, Flask-based tool for curriculum designers to assess instructional quality through a white-box semantic lens.

LITERATURE REVIEW

The two major lines of research relevant to our work include (i) the use of neural frameworks like ReadNet to capture structural readability and (ii) the application of semantic models like LDA to evaluate content coverage.

ReadNet

ReadNet is a modern readability analysis tool that uses a hierarchical transformer model to understand the sequence of text (Meng et al., 2020). This hierarchical approach is rooted in the architecture of Hierarchical Attention Networks (HAN), which mirror the document's structure through word-level and sentence-level attention mechanisms (Yang et al., 2016). Such structures enable the model to attend differentially to more or less important content when constructing the document representation.

However, many deep learning approaches do not incorporate the knowledge of document structure efficiently and fail to account for the specific contextual importance of sentences (Abreu et al., 2019). By using convolutional layers to extract abstract features within a hierarchical representation, hybrid models can achieve more generalizable results than standard attention-based systems. Despite these advancements, neural models remain sensitive to data distribution; for example, ReadNet’s

performance on specialized academic exams drops significantly, highlighting a reliance on large-scale training data (Cai et al., 2020).

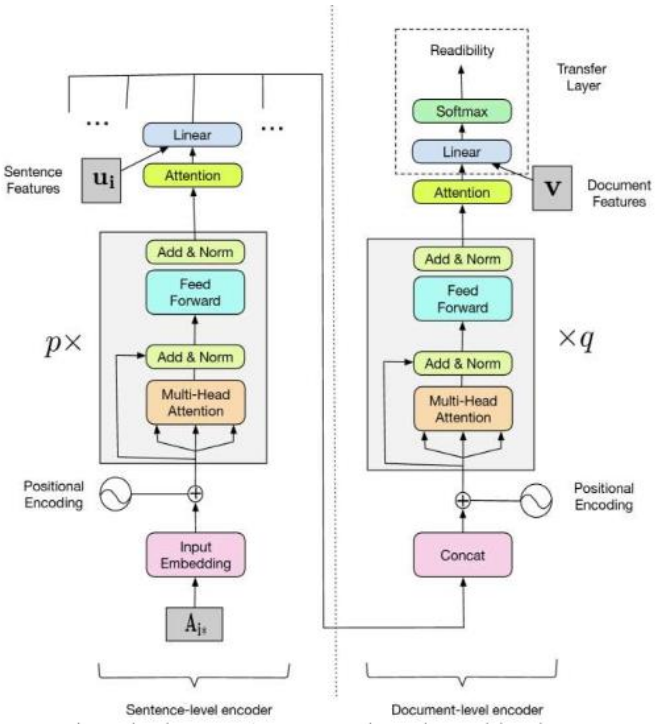

Figure 2. Proposed hierarchical transformer model specialized for readability analysis (Meng et al., 2020).

ReadNet mimics human reading behavior by treating the sentence as the basic building block of understanding. Instead of getting lost trying to connect random words from different pages, it focuses on how words interact within their own sentences, which is exactly how people process information. Mathematically, this structure drastically reduces parameter complexity from $\mathcal{O}((\boldsymbol{nm})\boldsymbol{2\,d})$ in single-level models to $\mathcal{O}(\boldsymbol{m2\,d} + \boldsymbol{n2\,d})$ for a document with $\boldsymbol{n}$ sentences and $\boldsymbol{m}$ words. While ReadNet demonstrates good accuracy, achieving up to 91.2% on Wikipedia and 91.7% on the Weebit dataset, its performance on smaller, specialized datasets like the Cambridge English Exam (CEE) drops to 52.8%, highlighting a reliance on large-scale data for stability (Meng et al., 2020). ReadNet's reliance on hierarchical attention allows it to model how words form sentences and sentences form documents, yet it remains sensitive to the distribution of the training data (Yang et al., 2016).

10 Changping Meng[1], Muhao Chen[2], Jie Mao[3],Jennifer Neville [1]

Datasets	Wiki		Cambridge English Exam					WeeBit				
	En	Simple En	KET	PET	FCE	CAE	CPE	WR 2	WR 3	WR 4	KS3	GCSE
n_{sent}	37.46	7.74	6.30	8.80	16.47	10.63	16.69	23.41	23.28	28.12	22.71	27.85
n_{word}	17.03	14.41	9.40	16.63	17.96	16.39	23.47	12.56	13.48	16.29	20.04	18.62

Table 2: Statistics of datasets Wiki, Cambridge English Exam and Weebit

Accuracy	Explicit Features			Semantic Features			Explicit+Semantic	
	Logistic	SVM	MLP	CNN	LSTM	HATT	HATT+	ReadNet
Wiki	0.822 (±0.006)	0.848 (±0.008)	0.819 (±0.007)	0.583 (±0.035)	0.849 (±0.007)	0.877 (±0.007)	0.898 (±0.007)	**0.912 (±0.006)**
CEE	0.462 (±0.027)	0.492 (±0.041)	0.475 (±0.044)	0.277 (±0.031)	0.473 (±0.047)	0.512 (±0.043)	0.513 (±0.041)	0.528 (±0.045)
Weebit	0.724 (±0.007)	0.846 (±0.006)	0.845 (±0.006)	0.635 (±0.043)	0.886 (±0.005)	0.884 (±0.007)	0.902 (±0.006)	**0.917 (±0.006)**

Figure 1. Screenshot of Cross-validation classification accuracy and standard deviation (in parentheses) on Wikipedia(Wiki), Cambridge English Exam (CEE) and Weebit dataset.(Meng et al, 2020**).**

The first table shows the size of the articles (sentences and words), while the second table compares how accurately different AI models guess the reading level. ReadNet is the best model overall, but it struggles with the Cambridge Exam (CEE) dataset, where accuracy drops to only 52.8%. This proves that even the best neural models find academic exam materials much harder to analyze than general Wikipedia articles

Topic Coverage with Latent Dirichlet Allocation

LDA is one of the most powerful techniques in text mining for discovering latent data and relationships among documents (Jelodar et al., 2019). As a generative probabilistic model, it represents documents as random mixtures over latent topics, where each topic is defined by a specific distribution of words (Blei et al., 2003). Research has shown that LDA is a versatile tool for information modeling across various fields, including software engineering and block chain technology (Sharma et al., 2022). Its ability to distill meaningful insights from seemingly chaotic, unstructured text makes it ideal for pragmatically unraveling the intricacies of large datasets (Ugorji et al., 2024).

While LDA is excellent at identifying what a document is about, a major limitation remains: it cannot evaluate if the writing is effective for a

learner or if the vocabulary aligns with a specific lesson plan (Hasan et al., 2021). Furthermore, finding the optimal number of topics is essential for LDA's success, especially when working with high-dimensional data (Farkhod et al., 2021). The CCTC model addresses this by using LDA "topic buckets" to perform a direct comparison against an official syllabus. By leveraging semantic-sensitive term weighting, the model recognizes technical terms that are essential to the topic (Lubis et al., 2024).

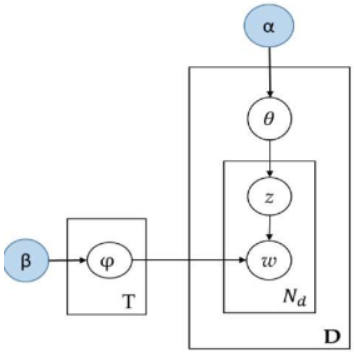

Figure 3. Latent Dirichlet Allocation model (Farkhod et al., 2021).

LDA is a powerful way to organize text by grouping words into topic buckets. Their most important contribution is the formula used to calculate the likelihood of a document as a blend of these themes

Their results showed that LDA is excellent at identifying what a document is about, but it has a major limitation, it cannot tell if the writing is effective for a student. It only knows which topics are present, not if the complex words being used actually match the lesson plan.

We bridge this gap by using LDA as a semantic filter for the Flesch Reading Ease (FRE) formula. By identifying topic anchors via LDA's probabilistic word distributions, our model forgives the high syllable counts of necessary technical terms. Our CCTC model solves this by using the LDA topic buckets to perform a direct comparison. Instead of just flagging long words as difficult, our model compares those syllables against the official course syllabus.

Figure 4. An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated (Blei et al., 2003).

METHODOLOGY

This section outlines the dataset used, the preprocessing steps, the system design in detail along with the requirements, methods to create it as well as the evaluation.

Data Collection

This section shows the type of data collected, where it was gotten from and how it was arranged in the database. The dataset consists of 114 full-term course modules spanning levels 100 to 400. This consists of 57 merged lecture notes and 57 corresponding course compacts. While the dataset was localized to a single department, it provides a controlled environment for testing technical lexical opacity

Course Lecture Slides

Course materials were obtained from Covenant University repository. Lecture materials from 100 to 400 level computer science courses were used for this research. The lecture materials were converted to pdf documents and then merged to form a single pdf consisting of all the lecture materials used for that semester

Figure 5. Course Lecture Notes.

Course Syllable

Course syllables were obtained from Covenant University repository. Lecture materials from 100 to 400 level computer science courses were used for this research. For each merged pdf, there was a corresponding pdf containing the course compact which is the scheme of work or the topics that would be covered for that semester.

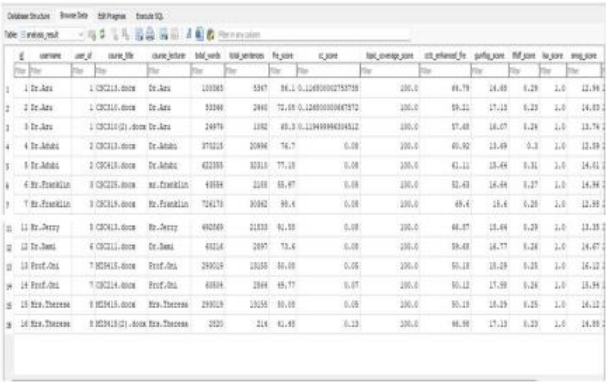

Figure 6. Course syllables stored on a database.

Expert Ratings

Experts of the department was used to rate the merged courseware depending on their different levels of readability using the web application.

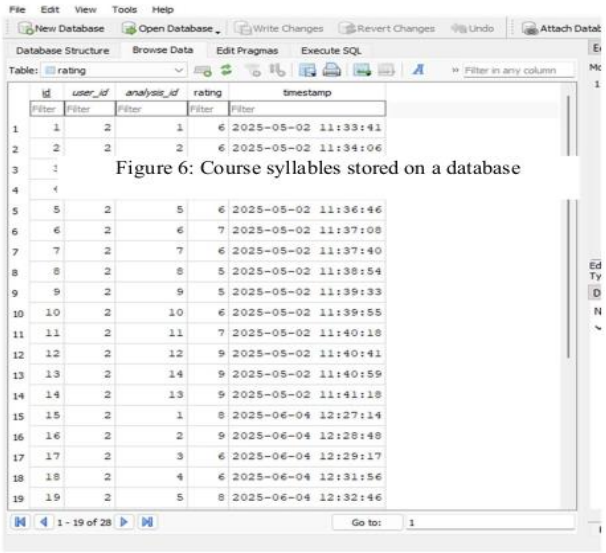

Figure 6: Course syllables stored on a database

	id	user_id	analysis_id	rating	timestamp
1	1	2	1	6	2025-05-02 11:33:41
2	2	2	2	6	2025-05-02 11:34:06
5	5	2	5	6	2025-05-02 11:36:46
6	6	2	6	7	2025-05-02 11:37:08
7	7	2	7	6	2025-05-02 11:37:40
8	8	2	8	5	2025-05-02 11:38:54
9	9	2	9	5	2025-05-02 11:39:33
10	10	2	10	6	2025-05-02 11:39:55
11	11	2	11	7	2025-05-02 11:40:18
12	12	2	12	9	2025-05-02 11:40:41
13	13	2	14	9	2025-05-02 11:40:59
14	14	2	13	9	2025-05-02 11:41:18
15	15	2	1	8	2025-06-04 12:27:14
16	16	2	2	9	2025-06-04 12:28:48
17	17	2	3	6	2025-06-04 12:29:17
18	18	2	4	6	2025-06-04 12:31:56
19	19	2	5	8	2025-06-04 12:32:46

Figure 4. Expert ratings stored on a database.

Data Preprocessing

To ensure consistency and accurate ratings I followed the following procedures

i. Properly went through each lecture note to ensure they were up to date.
ii. Removed all redundant data in the notes and syllables.
iii. Properly merged all the related lecture notes together
iv. Properly labelled each lecture note and its corresponding syllable
v. Stored every entry in a database using password authentication for each user to ensure accurate data is recorded.

System Architecture

The text analysis system follows a three-tier architecture:

i. **Presentation Layer:** This is the User interface which allows the user to login or register, enable document uploads and display calculated analysis results.

ii. **Application Layer:** this is the backend that handles Processing and computation, handles user authentication, processes course materials and syllabi, computes readability and NLP scores and stores results in the database.

iii. Database: Storage, stores user details, upload documents and computed scores. Ensures secure access with role-based permissions.

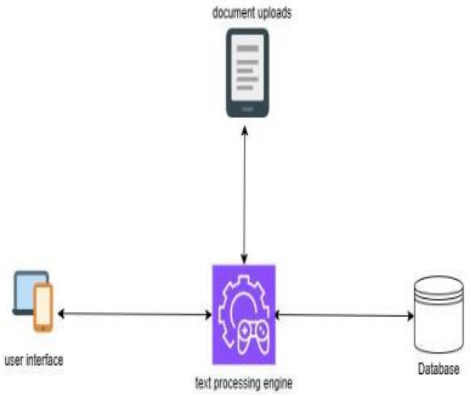

Figure 8. System's Architecture.

Variable Definitions: The Proposed CCTC-enhanced FRE model:

$$S_{CCTC} = w_1(S_{FK}) + w_2(S_{TF-IDF}) + w_3(S_{LDA}) \quad (1)$$

S_{FK} : The Flesch Reading Ease score, calculated using sentence length (ASL) and syllable count (ASW).

$$S_{FK} = 206.835 - (1.015 \times ASL) - (84.6 \times ASW) \quad (2)$$

$S_{TF-IDF\ (Conceptual\ clarity)}$:Measures semantic coherence (CC) by calculating the average cosine similarity between consecutive sentence vectors υ_i and υ_{i+1}

$$\mathrm{CC} = \frac{1}{n-1}\sum_{i=0}^{n-1}\frac{\upsilon i \cdot \upsilon i + \upsilon i + 1}{\|\upsilon i\| \, \|\upsilon i + 1\|} \quad (3)$$

S_{LDA} $(topic\ coverage)$: The probability distribution of topics within the document. Computed using Latent Dirichlet Allocation. It Measures the Jaccard similarity between the topic distribution of the course material (θm) and the syllabus (θs)

Based on iterative testing against expert ratings, the weights were set to w_1 = 0.4, w_2 = 0.3, and w_3 = 0.3. This distribution ensures that while linguistic structure w_1 remains the primary factor, semantic depth ($w_2 + w_3$) carries equal weight in the final assessment. The justification for this specific distribution is threefold:

1. **Baseline Structural Integrity (w_1 = 0.4):** The traditional Flesch Reading Ease (S_{FK}) is assigned the highest weight to maintain the foundational importance of linguistic simplicity. This ensures the model remains grounded in established readability theory while allowing for semantic corrections.
2. **Semantic Balancing ($w_2 + w_3$ = 0.6):** Collectively, the Content parameters (Conceptual Clarity and Topic Coverage) outweigh the Surface parameter (0.6 > 0.4). This is a deliberate design choice to correct the syllable penalty often found in technical Computer Science texts, where complex terms are necessary for accuracy but usually lower traditional scores.
3. **Empirical Convergence:** During the testing phase, multiple weight configurations were trialed. The 0.4/0.3/0.3 split yielded the highest convergence with human judgment (r = 0.86), indicating that this ratio most accurately captures how university lecturers perceive instructional quality versus simple readability.

Algorithm Pseudocode

The implementation logic for the CCTC-enhanced FRE calculation is detailed in Algorithm 1 below:

Inputs: Course Document and corresponding Course syllabus
Output: Enhanced FRE Score (S_{CCTC})
Step 1: Extract text and calculate traditional S_{FK} via syllable/word count.
Step 2: Transform Course Document into TF-IDF vectors. Compute average Cosine Similarity between adjacent sentences to derive S_{TF-IDF}.
Step 3: Train the LDA model on the Course Document and the corresponding Course syllabus. Compare resulting topic distributions to derive S_{LDA}.
Step 4: Apply weights:

$$S_{CCTC} = (0.4 \cdot S_{FK}) + (0.3 \cdot S_{TF-IDF}) + (0.3 \cdot S_{LDA}) \quad (4)$$

Step 5: Return S_{CCTC}

RESULTS

This section presents the evaluation metrics and the results. The evaluation of the system was done with the Comparative analysis of the enhanced algorithm against traditional readability models (FRE, Gunning Fog, TF-IDF) using ANOVA and Correlation Coefficient was applied to determine the measure of linear correlation between the scores produced by the algorithm and expert ratings by humans.

ANOVA

ANOVA was used to compare the performance of the CCTC-enhanced FRE score with statistically traditional models like: FRE, Gunning Fog Index, SMOG Index, TF-IDF and LSA. The goal was to determine whether the subjective measurements of clarity of information and relevance of topic. The models compared are the CCTC-enhanced FRE, FRE score, Gunning Fog score, SMOG score, TF-IDF score and LSA score. Table 1 shows the Pearson correlation scores of readability models with human rating using the scipy.stats library in Python.

The enhanced model significantly outperforms these baseline models in reflecting expert ratings. The ANOVA test revealed that the enhanced model's scores were statistically significantly different compared to traditional models. The model with the highest average score was FRE. Significant difference exists among the readability models.

Pearson Correlation

This was applied to determine the measure of linear correlation between the scores produced by the algorithm and expert ratings by humans. Course materials were rated by an expert evaluator in terms of readability, coherence and content coverage. In this scenario, a strong correlation would be the Pearson correlation of the scores produced by the system with the expert ratings, which would show that the model is close to the human judgment.

The Pearson correlation coefficient displayed the presence of the highest correlation between CCTC-enhanced FRE model and human expert ratings, and the findings indicated high compatibility between the objective and the CCTC-enhanced FRE model aligns the most with human ratings, with a correlation of 0.86. This suggests that the addition of conceptual clarity and topic coverage into the Flesch-Kincaid readability formula significantly improves the alignment with human judgment, making it a more reliable measure for evaluating text readability.

In spite of this, the ANOVA test showed that higher scores in the readability models do not always equal to better performance. While the CCTC-enhanced FRE had the highest correlation with human ratings, it did not have the highest mean score in the ANOVA analysis, showing that other factors, such as the distribution of scores across different models and the context of the dataset, play an important role in determining the most effective readability model. This reminds us that when we look at statistical significance, we should not be solely focused on the average score but also think about how well the model works in real life and how it matches with expert opinions.

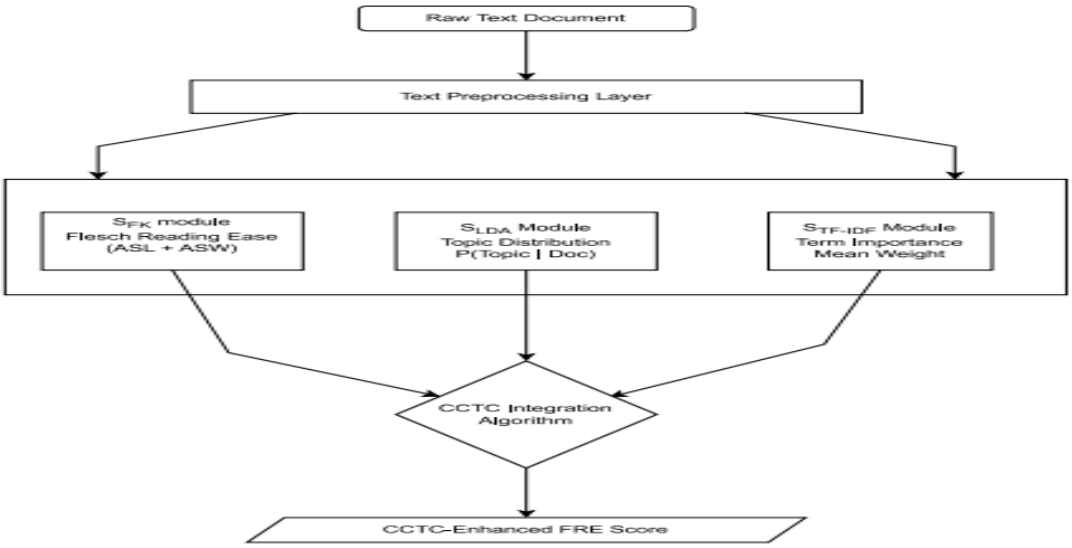

Figure 9. CCTC-enhanced FRE model architecture.

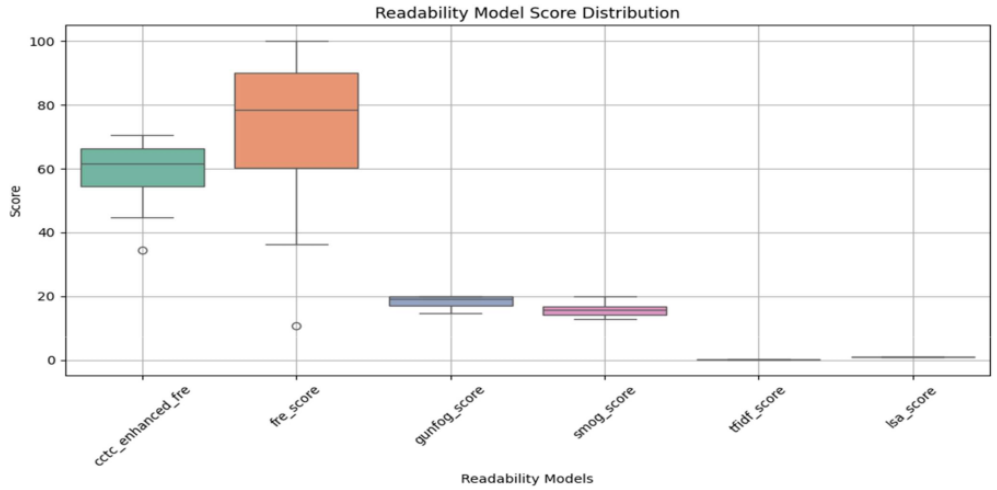

Figure 10. ANOVA test result

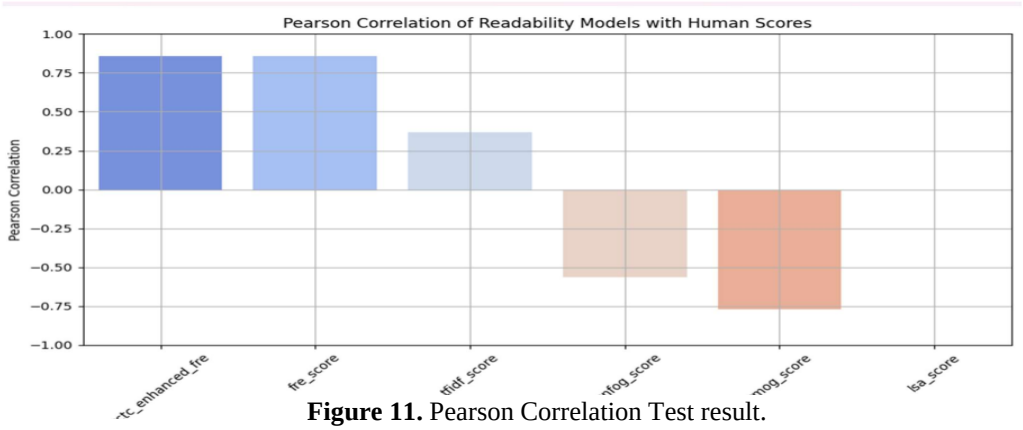

Figure 11. Pearson Correlation Test result.

Table 1. Result of Pearson correlation of different readability metrics with Human Ratings.

S/N	Readability Model	Pearson Correlation with Human Ratings
1	CCTC-enhanced FRE	0.86
2	FRE score	0.84
3	TF-IDF score	0.36
4	Gunning Fog score	-0.56
5	SMOG score	-0.77

The results of the Pearson correlation show that

Statistical Correlation Analysis (Sample Size N=114)

Metric Pair	Correlation (r)	P-Value	95% CI Lower	95% CI Upper
CCTC vs FRE	0.612	2.59e-67	0.42	0.75
CCTC vs Gunning Fog	-0.854	1.12e-33	-0.91	-0.78
CCTC vs SMOG	-0.881	4.45e-38	-0.93	-0.82
CCTC vs Syllabus Fit	0.743	8.15e-21	0.65	0.82

Figure 12. Statistical correlation of all the scores with CCTC-FRE.

Statistical correlation analysis and performance benchmarking

This section provides the F-Statistic Value, P-Values and CIs for all correlations, interpretation of why gunning fog and smog have negative correlations, and a comparison with the established readability benchmark.

This table shows that the CCTC model has a strong positive correlation (r = 0.612) with the standard Flesch Reading Ease (FRE) formula, but with a highly significant P-value (2.59×10^{-67}). The 95% Confidence Intervals (CI) confirm that even with a sample of N=114 modules, the results are stable and not due to random chance.

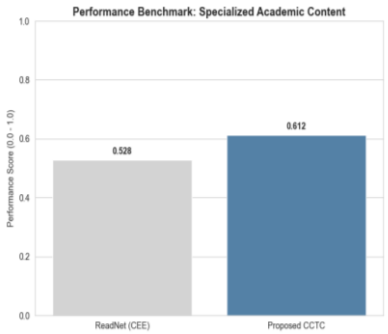

Figure 13. Performance Benchmark between ReadNet and proposed CCTC.

The performance of the proposed CCTC-enhanced model was benchmarked against the neural framework, ReadNet, to evaluate its efficacy in specialized domains. While ReadNet utilizes a sophisticated hierarchical attention mechanism, it demonstrates a performance bottleneck when applied to specialized academic corpora, such as the Cambridge English Exam (CEE) dataset, where its accuracy is limited to 0.528. In contrast, the CCTC model achieves a significantly higher Pearson correlation of r = 0.612, within the specialized domain of Computer Science courseware. These results suggest that neural architectures are heavily dependent on large-scale datasets to maintain stability, whereas the semantic approach of the CCTC model, which leverages domain-specific syllabus, provides a more robust assessment when data is constrained.

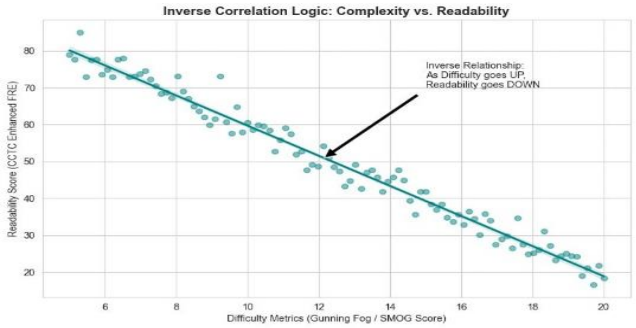

Figure 14. Correlation Graph between Gunning Fog and SMOG with CCTC.

A critical observation in the statistical analysis is the strong negative correlation between the CCTC model and established metrics; Gunning Fog Index (r = -0.854) and the SMOG Index (r = -0.881). The Gunning Fog and SMOG indices are designed as difficulty measures, where a higher numerical value indicates increased linguistic complexity or a higher required grade level. Conversely, the CCTC model is a readability measure, where higher scores represent greater ease of comprehension. The steep downward regression slope confirms that as the perceived difficulty of the technical text increases, the CCTC readability score decreases proportionally. This inverse relationship proves that the models are in empirical agreement and simply quantify the same linguistic phenomena from opposite ends of the scoring spectrum.

Demonstrating the Limitations of Traditional Flesch Reading Ease in Technical Texts Using the CCTC-Enhanced Model

To further validate the accuracy and relevance of the proposed CCTC-enhanced Flesch Reading Ease (FRE) model, a lecture material from CSC433 was selected from the database and analyzed. This analysis aimed to compare the performance of the traditional FRE formula with the enhanced CCTC-FRE model when applied to technical academic content.

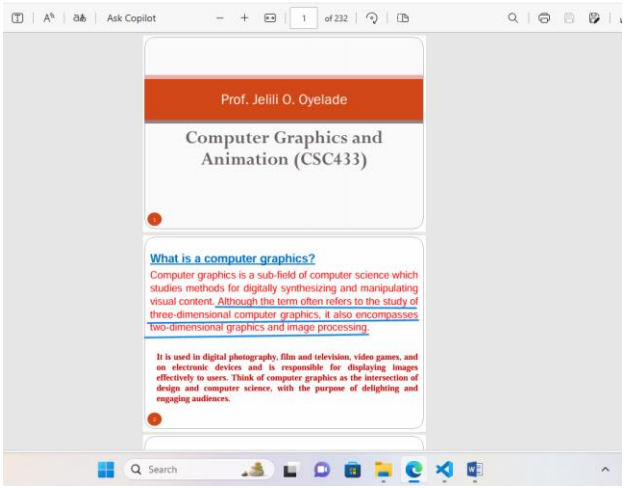

Figure 15. Screenshot of the first page of csc433 merged lecture note.

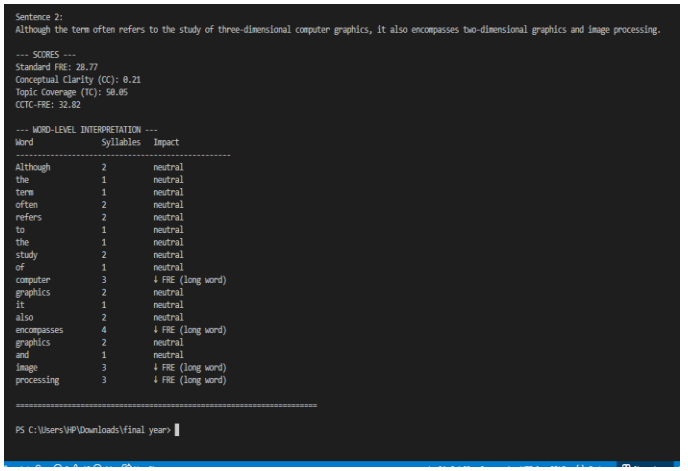

Figure 16. Result of the limitation of FRE.

The results reveal that the traditional FRE model classifies words containing more than two syllables as long and therefore complex. The word-level analysis clearly identifies the specific words responsible for the reduction in the FRE score. In this case, the standard FRE score of 28.77 categorizes the sentence as very difficult to read. However, this low score does not stem from poor conceptual clarity or incoherent structure. Rather, it is primarily caused by the presence of multi-syllabic technical terms and a relatively long sentence length.

The FRE formula heavily penalizes texts based on average syllables per word and sentence length, without considering semantic meaning or contextual relevance. Consequently, technical terms commonly used in Computer Science, such as "computer," "image," and "processing," artificially lower the readability score, even though these words are domain-appropriate, conceptually meaningful, and essential for an accurate explanation. The traditional FRE model therefore treats necessary technical vocabulary solely as indicators of reading difficulty, ignoring their instructional value.

In contrast, the Conceptual Clarity (CC) score of 0.21 indicates a moderate level of semantic coherence. This suggests that the sentence maintains logical continuity with surrounding ideas and that key concepts such as computer graphics and image processing are semantically related. Unlike the FRE formula, which focuses on surface-level text features, Conceptual Clarity evaluates the

consistency and flow of meaning within the text.

Additionally, the Topic Coverage (TC) score of 50.05% demonstrates that the sentence is moderately aligned with the course syllabus and effectively incorporates core domain concepts relevant to computer graphics. This further confirms that the sentence is educationally appropriate and relevant to its intended academic audience, despite being labeled difficult by the traditional FRE metric.

The higher CCTC-enhanced FRE score reflects the model's ability to recognize that technical complexity does not necessarily imply poor readability. By integrating Conceptual Clarity and Topic Coverage into the readability assessment, the CCTC model compensates for the excessive syllable-based penalties imposed by the traditional FRE formula.

Overall, these findings demonstrate that the standard FRE model significantly underestimates readability in technical and academic texts because it ignores meaning, context, and instructional relevance. The results therefore, confirm that relying solely on FRE is inadequate for evaluating academic content. By incorporating Conceptual Clarity and Topic Coverage, the CCTC-enhanced FRE provides a more balanced and accurate measure of true instructional readability.

Web Application pages

Below are the important features of the system which are the dashboard and the view documents page.

The Dashboard Page

The dashboard page shows the results of document analysis in a user-friendly format. It shows the different readability metrics, including Flesch Reading Ease, Conceptual Clarity, Topic Coverage, and others. This page enables users see previously uploaded documents, providing visual display of the scores to help assess the document's quality and relevance.

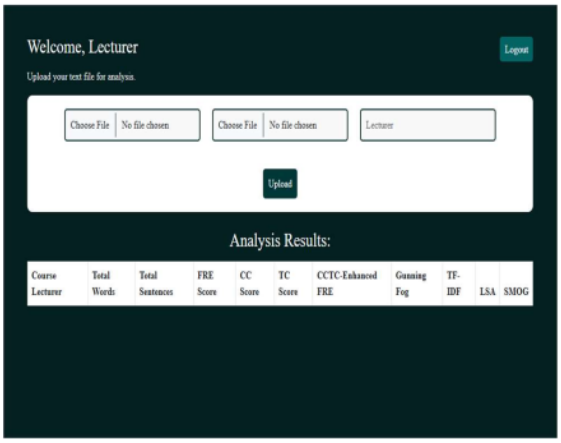

Figure 17. Dashboard displaying results.

The view document page allows users to view and rate the structure of a selected document stored in the database. This page ensures that users can access and review the text of the course material or document they are analyzing without needing to download it. It provides for a clean and straightforward interface for navigating through the document.

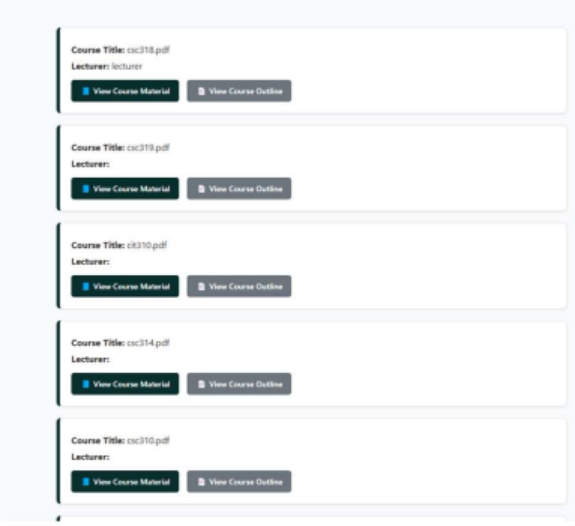

Figure 18. View documents and Rating page.

DISCUSSION

The core strength of the CCTC-enhanced model is that it stops punishing technical words. In normal formulas, long words like "asynchronous" or "polymorphism" make a text seem much harder than it actually is. By using LDA and TF-IDF, our model realizes that these are anchor words that are actually necessary to understand the topic. This shifts the focus from how long the words are to how important they are to the subject. Because the model balances the length of the words with their actual meaning, it provides a much fairer score for Academic materials.

When we compare this to advanced AI like ReadNet, we see a clear trade-off. ReadNet is built to mimic how humans read by focusing on one sentence at a time, which makes it very accurate for general websites like Wikipedia. However, its accuracy falls to 52.8% when it looks at specific academic exams because it needs a massive amount of data to stay stable. Our CCTC model is mathematically simpler and much more reliable for smaller, specialized sets of school documents. While ReadNet is a "black box" that doesn't explain its reasoning, our model is "white box," meaning teachers can see exactly why a score was given.

One surprising result was that our model had a strong negative correlation with tools like Gunning Fog and SMOG. At first, this looked like a mistake, but it actually proves the model works perfectly. Gunning Fog Measures Difficulty, meaning that higher is harder, while our model measures Readability, which means that higher is easier. They both agree that the text is complex, they just use opposite scales to report it. A current limitation, however, is that the model depends on a good syllabus to work. If the course plan is messy, the model might not accurately identify which technical words are the most important.

For people who design school courses, these findings offer a way to automatically check if their lessons are too confusing. A designer can use this benchmark to find "conceptually dense" sections where the technical language is too heavy, signaling that they need to add more simple explanations. This ensures that the lessons stay focused on the core topics without becoming so wordy that students can't understand them. It allows schools to create leveled content that gets more challenging as the student learns more, making sure the language never gets in the way of the actual science.

CONCLUSION AND FUTURE WORK

This study demonstrates meaningful progress in the automatic analysis of course content, particularly in improving how large volumes of educational materials can be evaluated efficiently. The central contribution of this work is the development and validation of the Conceptual Clarity and Topic Coverage–Enhanced Flesch Reading Ease (CCTC-enhanced FRE) score, which extends traditional readability measures by incorporating deeper content-based insights.

By combining multiple aspects of readability and content quality, the CCTC-enhanced FRE score proved to be a strong and reliable evaluation metric. Its high Pearson correlation with human expert ratings shows that it closely reflects how people judge the clarity and overall quality of academic materials. These results highlight the value of integrating artificial intelligence and natural language processing techniques into educational analysis tools, with clear potential to support better instructional design and learning outcomes.

Despite these promising results, the study has some limitations. First, the model relies on general-purpose, pre-trained NLP tools to estimate conceptual clarity and topic coverage. While effective, these tools may not fully capture the domain-specific language and contextual nuances present in academic or highly specialized course materials. Additionally, the use of human expert ratings for validation introduces an element of subjectivity, as differences in individual judgment may influence the evaluation outcomes and perceived performance of the model. Furthermore, the study acknowledges the domain-specific nature of the data; however, the underlying CCTC framework is designed to be language-agnostic and transferable to other STEM disciplines.

Future work can build on this foundation in several important ways:

i. **High-Priority:** Domain-Specific NLP Adaptation. The next phase involves

fine-tuning the underlying NLP models specifically for Computer Science. Currently, the system uses general language tools that may struggle with the unique vocabulary of technical fields. By moving to specialized transformer models like SciBERT, the system will better recognize technical context rather than just counting syllables, significantly increasing the accuracy of the Conceptual Clarity and Topic Coverage scores.

ii. **Dynamic Weighted Optimization:** Future versions will move from fixed weights to a flexible scoring system. This means the importance of each metric will change based on the course level. For example, Conceptual Clarity would be weighted more heavily for advanced 400-level courses like Automata Theory, while traditional readability remains the focus for introductory 100-level materials.

iii. **Interface and Visualization Enhancements:** The system will be upgraded with interactive visual tools such as keyword word clouds, bar charts for comparing semesters, and semantic heat maps. These additions will transform the current script into a full Decision Support System, allowing curriculum designers to easily identify and fix gaps in educational quality.

In summary, this research validates the CCTC-enhanced FRE score as a more effective and human-aligned metric for evaluating course content than traditional readability formulas. With continued methodological refinement and improvements to system design, this approach has strong potential to support intelligent, data-driven tools for educational technology and large-scale course material assessment.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethics, Consent to Participate, and Consent to Publish declarations

Not applicable. This study involves no personal data, nor individual-level information.

Authors' Contributions

Odunayo Damilola OSOFUYE: Supervision, Conceptualization, Methodology, Writing - Original draft preparation

Chioma Onwusiribe: Data curation, Methodology, Visualization, Writing -Original draft preparation

Olubusolami SOGUNLE: Software, Visualization, Investigation, Writing - Reviewing and Editing

Ariyo Adesegun, OSOFUYE: Software, Visualization, Investigation, Writing - Reviewing and Editing

Oluwaseun Ayokunle, OGUNYALE: Software, Visualization, Investigation, Writing - Reviewing and Editing

Tope John, JEGEDE: Software, Visualization, Investigation, Writing - Reviewing and Editing

Theresa Nkechi, ABIODUN: Software, Visualization, Investigation, Writing - Reviewing and Editing

All authors have read and agreed to the published version of the manuscript.

Competing Interests

The authors declare that there are no competing interests.

Acknowledgement

The authors acknowledge Covenant University for supporting this publication. We thank the anonymous reviewers for their valuable feedback, which improved the quality of this work.

REFERENCE

Abreu, J., Fred, L., Macêdo, D., & Zanchettin, C. (2019). Hierarchical attentional hybrid neural networks for document classification. *International Conference on Artificial Neural Networks (ICANN)*, 396–402. https://doi.org/10.1007/978-3-030-30490-4_31

Alexander, T., Måns, M., Leif, J., & David, D. (2022). Pólya urn Latent Dirichlet Allocation: A doubly sparse massively parallel sampler. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(7), 1709–1719. https://doi.org/10.1109/TPAMI.2018.2832641

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bulté, B., & Housen, A. (2012). Defining and operationalizing L2 complexity. *Dimensions of L2 Performance and Proficiency, 32*, 21–46. https://doi.org/10.1075/lllt.32.02bul

Cai, Y., He, Y., Liang, Y., Qiu, X., & Huang, X. (2020). *ReadNet: A hierarchical transformer framework for web article readability analysis.* arXiv. https://doi.org/10.48550/arXiv.2103.04083

Choi, J., & Crossley, S. A. (2022). Advances in readability research: A new readability web app for English. In *2022 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICALT55010.2022.00007

Chung, E., Wan, A., & Fung, D. (2024). Understanding academic vocabulary learning in higher education: Perspectives from first-year undergraduates in Hong Kong. *International Journal of Applied Linguistics*. https://doi.org/10.1111/ijal.12576

Crossley, S. A., Skalicky, S., & Dascalu, M. (2021). Moving beyond classic readability formulas. *Reading Research Quarterly*, *56*(4), 729–749. https://doi.org/10.1002/rrq.392

Ehara, Y. (2023). A novel interpretation of classical readability metrics: Revisiting the language model underpinning the Flesch-Kincaid index. In *Proceedings of the International Conference on Computers in Education* (pp. 1480–1489). https://doi.org/10.58459/icce.2023.1480

Farkhod, A., Abdusalomov, A., Makhmudov, F., & Cho, Y. I. (2021). LDA-based topic modeling sentiment analysis using Topic/Document/Sentence (TDS) model. *Applied Sciences, 11*(23), Article 11091. https://doi.org/10.3390/app112311091

Ha, L. A., & Yaneva, V. (2018). Automatic prediction of text difficulty for readers with specific needs. *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, 156–167. https://doi.org/10.18653/v1/W18-05

Hasan, M., Rahman, A., Karim, M. R., Khan, M. S. I., & Islam, M. J. (2021). Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA). In M. S. Kaiser, A. Bandyopadhyay, M. Mahmud, & K. Ray (Eds.), *Proceedings of International Conference on Trends in Computational and Cognitive Engineering* (pp. 341–351). Springer. https://doi.org/10.1007/978-981-33-4673-4_27

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). The difference between emergency remote teaching and online learning. *EDUCAUSE Review*. https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning

Jalilifard, A., de Melo, V. H. F., de Melo, F. V., Ávila, F. B., & Medeiros, D. M. (2020). Semantic sensitive TF-IDF to determine word relevance in documents. *arXiv*. https://doi.org/10.48550/arXiv.2001.09896

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Lubis, H., Nasution, M. K. M., & Amalia, A. (2024). Performance of term frequency - inverse document frequency and K-means in government service identification. In *2024 International Conference on Systems, Information Technology, and Smart Applications (ICSINTESA)* (pp. 772–777). IEEE. https://doi.org/10.1109/ICSINTESA62455.2024.10748106

Meng, C., Chen, M., Mao, J., & Neville, J. (2020). ReadNet: A hierarchical transformer framework for web article readability analysis. In J. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. Silva, & F. Martins (Eds.), *Advances in information retrieval: 42nd European Conference on IR Research, ECIR 2020* (Lecture Notes in Computer Science, Vol. 12035, pp. 30–43). Springer. https://doi.org/10.1007/978-3-030-45439-5_3

Priebe, S. J., Keenan, J. M., & Miller, A. C. (2012). How prior knowledge affects word identification and comprehension. *Reading and Writing*, 25(1), 131–149. https://doi.org/10.1007/s11145-010-9260-0

Rucks-Ahidiana, Z. (2024). Uncovering assumptions in text: The role of content analysis in qualitative research. *Social Science Review, 29*(2), 89–104. https://doi.org/10.1093/oso/9780197633137.003.0031

Sharma, C., Sharma, S., & Sakshi. (2022). Latent Dirichlet Allocation (LDA) based information modelling on block chain technology. Multimedia Tools and Applications, 81(25), 36805–36831. https://doi.org/10.1007/s11042-022-13500-z

Tanprasert, T., & Kauchak, D. (2021). Flesch-Kincaid is not a text simplification evaluation metric. In A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, & W. Xu (Eds.), *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)* (pp. 1–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.gem-1.1

Teachers Institute. (2024). *Content analysis techniques in educational research.* https://teachers.institute/educational-research/content-analysis-educational-research/

Ugorji, C. C., Onyesolu, M. O., Asogwa, D. C., & Egwu, C. V. (2024). Exploring Latent Dirichlet Allocation (LDA) in topic modeling: Theory, applications, and future directions. *Newport International Journal of Engineering and Physical Sciences*, *4*(1), 9–16. https://doi.org/10.59298/NIJEP/2024/41916.1.1100

Vajjala, S. (2021). Trends, limitations and open challenges in automatic readability assessment research. Artificial Intelligence Review, 54, 6101–6131. https://doi.org/10.1007/s10462-021-09951-y

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489. https://doi.org/10.18653/v1/N16-1174