http://dx.doi.org/10.18623/rvd.v22.n4.3612

REDUNDANCY REDUCTION AND SENTENCE PRIORITISATION OF THE STUDENT LECTURE NOTES USING SOFT COSINE IMPLEMENTED MMR ALGORITHM

REDUÇÃO DE REDUNDÂNCIA E PRIORIZAÇÃO DE FRASES NAS NOTAS DE AULA DO ALUNO USANDO O ALGORITMO MMR IMPLEMENTADO COM COSSENO SUAVE

Article received on: 7/21/2025
Article accepted on: 10/27/2025

Angelo Baby*
*Christ University, Bangalore, Rajagiri College of Social Sciences, Kerala, Kerala, India
Orcid: https://orcid.org/0000-0002-8822-2825
angelo.baby@res.christuniversity.in

Vaidehi V**
**Christ University, Bangalore, Karnataka, India
Orcid: https://orcid.org/0000-0002-2293-4934
vaidehi.v@christuniversity.in

Jinsi Jose***
***St. Joseph's College Moolamattom (Autonomous), Idukki, India
Orcid: https://orcid.org/0000-0002-4691-6474
jinsi.jose@stjosephscollegemoolmattom.ac.in

The authors declare that there is no conflict of interest

Abstract
In the realm of education, the efficacy of lecture notes in aiding student learning is pivotal. This study explores the integration of the Soft cosine measure (SCM) with the Maximal marginal relevance (MMR) algorithm to reduce redundancy and prioritize essential sentences in student lecture notes. Traditional cosine similarity often overlooks the semantic similarity between terms with different surface forms, leading to suboptimal redundancy reduction. The Soft Cosine Measure addresses this by accounting for word similarity based on semantic relationships, improving sentence relevance and uniqueness assessment. In our approach, SCM is employed within the MMR framework to iteratively select sentences that maximize relevance to the main lecture content while minimizing redundancy with previously selected sentences. This hybrid method ensures that the final summary encompasses a broader range of key concepts and topics discussed during lectures, providing a more comprehensive and coherent overview. Experimental evaluations on a dataset of lecture notes demonstrate that the SCM-implemented MMR algorithm significantly outperforms traditional summarization techniques in both reducing redundancy and maintaining the informativeness

Resumo
No âmbito da educação, a eficácia das notas de aula no auxílio à aprendizagem dos alunos é fundamental. Este estudo explora a integração da Medida de Cosseno Suave (MCS) com o algoritmo de Relevância Marginal Máxima (RMM) para reduzir a redundância e priorizar frases essenciais nas notas de aula dos alunos. A similaridade de cosseno tradicional frequentemente ignora a similaridade semântica entre termos com diferentes formas superficiais, levando a uma redução de redundância subótima. A Medida de Cosseno Suave resolve esse problema ao considerar a similaridade entre palavras com base em relações semânticas, melhorando a relevância das frases e a avaliação da singularidade. Em nossa abordagem, a MCS é empregada dentro da estrutura RMM para selecionar iterativamente frases que maximizem a relevância para o conteúdo principal da aula, minimizando a redundância com frases previamente selecionadas. Esse método híbrido garante que o resumo final abranja uma gama mais ampla de conceitos e tópicos-chave discutidos durante as aulas, proporcionando uma visão geral mais abrangente e coerente. Avaliações experimentais em um conjunto de dados de notas de aula demonstram que o algoritmo MMR implementado com SCM supera significativamente as técnicas tradicionais de

of summaries. The Soft Cosine Implemented MMR Algorithm (SCIMMR) gives the ROUGE values between 0.821 to 0.901. This novel approach offers a robust solution for the automatic summarization of lecture notes, aiding students in efficiently reviewing and studying educational materials.

Keywords: Maximal Marginal Relevance (MMR). Redundancy Reduction. Sentence Prioritisation. Soft Cosine. Contextual Tokenization.

sumarização, tanto na redução da redundância quanto na manutenção da informatividade dos resumos. O Algoritmo MMR Implementado com Cosseno Suave (SCIMMR) apresenta valores ROUGE entre 0,821 e 0,901. Essa nova abordagem oferece uma solução robusta para a sumarização automática de notas de aula, auxiliando os alunos na revisão e no estudo eficiente de materiais didáticos.

Palavras-chave: *Relevância Marginal Máxima (MMR). Redução de Redundância. Priorização de sentenças. Cosseno Suave. Tokenização contextual.*

1 INTRODUCTION

In the digital era of education, where information overload is a common challenge, optimizing student lecture notes has become increasingly crucial. Efficient organization and prioritization of content facilitate comprehension and aid in knowledge retention. Integrating advanced computational techniques, particularly natural language processing (NLP) algorithms, offers promising solutions in this context. This research explores the implementation of the maximal marginal relevance (MMR) algorithm, enhanced with soft cosine similarity, to address the twin objectives of redundancy reduction and sentence prioritization in student lecture notes. This research article presents a comprehensive framework for redundancy reduction and sentence prioritization in student lecture notes using the Soft Cosine Implemented MMR Algorithm (SCIMMR) by using real time dataset.

Redundancy reduction and sentence prioritization emerge as critical strategies to alleviate cognitive overload and enhance the utility of lecture notes. Reducing redundancy ensures that essential information is conveyed concisely, minimizing unnecessary repetition and maximizing the clarity of content. Meanwhile, prioritizing sentences allows students to focus on key concepts and insights, facilitating deeper understanding and knowledge retention.

This research aims to explore the application of the MMR algorithm, augmented with soft cosine similarity, to optimize student lecture notes. MMR, initially proposed by Carbonell and Goldstein (Carbonell & Goldstein, 1998), has shown promise in information retrieval and summarization tasks, while soft cosine similarity, introduced by Sidorov *et al*. (Sidorov1 *et al*., 2014), extends traditional cosine similarity by incorporating semantic information, thereby providing a more nuanced measure of textual similarity. By integrating these techniques, we seek to develop a comprehensive framework for redundancy reduction and sentence prioritization in student lecture notes.

Redundancy reduction is a critical aspect of knowledge extraction, aimed at eliminating repetitive information from text documents to enhance clarity and conciseness. Traditional methods such as clustering and summarization have been widely employed for this purpose (Bucilă *et al*., 2006; Erkan & Radev, 2011). However, these approaches often overlook the semantic relationships among sentences, leading to suboptimal results. In contrast, the soft cosine similarity (SCS) metric computes the similarity between text documents while preserving semantic meaning (Leskovec *et al*., 2011). By leveraging SCS, our proposed framework can identify and eliminate redundant sentences with greater precision, thereby improving the coherence and relevance of distilled lecture notes.

Redundancy reduction techniques aim to identify and eliminate duplicate or semantically similar information from lecture notes (Locke, 2015). Various methods have been proposed to achieve this, including lexical analysis, semantic similarity assessment, and summarisation algorithms (Colombo, 2024). Soft Cosine Similarity, a measure of semantic similarity between text documents, has emerged as an effective tool for identifying redundant information in lecture notes (Jain & Rastogi, 2020). By quantifying the semantic relatedness between sentences, redundant content can be consolidated, resulting in more concise and informative notes (Mao *et al*., 2020). This approach not only reduces cognitive load for students but also enhances the overall clarity and coherence of the notes (Renkl & Atkinson, 2003).

Sentence prioritisation involves ranking sentences based on their relevance and importance to the lecture topic (Upadhay & Singh, 2020). Prioritisation algorithms, MMR offer a systematic framework for selecting the most informative sentences while minimising redundancy (Gunawan *et al*., 2023). MMR considers both the relevance and diversity of sentences, ensuring that the selected sentences capture the essence of the

presentation while avoiding redundancy (Prasetya & Kurniawan, 2024). By integrating MMR with redundancy reduction techniques, lecture notes can be optimised to highlight key concepts and insights, facilitating student comprehension and retention (Demilie, 2022). Subsequently, the MMR algorithm is employed to prioritise sentences based on their relevance and diversity, ensuring that the most informative content is retained in the final notes (Carbonell, 1998). Another approach (Aranzamendez, S.G, *et al*.,2024) adds MMR re-ranking to a content-based recommender to fight overspecialization (too-similar items). Tunes the MMR λ trade-off; reports higher precision/F1 while lifting diversity/novelty. Shows a simple way to diversify any content-based list without a full model rewrite. (Wang, Z., *et al*., 2024) represents papers via topic combinations (e.g., BER Topic) and defines novelty as "organic reorganization" of topics. Uses an MMR-based topic selection step to balance similarity vs. diversity when characterizing a paper and applies a cloud-model to score novelty; validates on real corpora (e.g., ICLR subset). They have proved MMR isn't just for search/recs useful in feature selection when novelty/diversity matter.

One commonly used metric in information retrieval and related fields is cosine similarity. This metric represents a text as a vector of terms, and the cosine value between the term vectors of two texts is used to calculate how similar two texts are to one another (Faisal *et al*., 2012). In this study (Faisal *et al*., 2012), a semantic verification between the dimensions of two term vectors is proposed as an improvement to the cosine similarity measurement. The objective of this approach is to raise the similarity value between two term vectors with distinct syntax that have semantic relationships between their dimensions. It is crucial to get precise results from the processing of documents, including text. The kind of analysis that needs to be done usually determines how approaches are implemented in relation to documents. When handling document similarity, the use of cosine similarity techniques is crucial, and many computer languages have different algorithms for implementing it (Januzaj & Luma, 2022). In (Jiang, P., & Cai, X., 2024), conducted survey statistical neural and LLM-era text matching. Explains where soft-cosine fits (captures inter-term similarity) vs. standard cosine (assumes orthogonal features). (Kaur, N., 2024) discussed about educational review of string/term-based similarities with a clear section on soft-cosine (how it relaxes the independence assumption). Good for definitions, formulas, and quick contrasts between cosine/soft-cosine/Jaccard/etc. In another study (Alizadeh, M., & Seilsepour, A.,

2025) builds a pseudo-label generator using soft-cosine similarity between samples and sentiment lexicons. Introduces SCSP/SCSN (soft-cosine to positive/negative word sets) to better handle domain-specific polarity. Improves sentiment accuracy without manual labels by leveraging semantic proximity rather than literal token overlap. (Ijebu, F.F., 2025), proposes adaptations of soft-cosine and extended-cosine geared to PLM/LLM embeddings (where feature overlap $\neq 0$ even when tokens differ). Benchmarks vs. standard cosine on semantic similarity style tasks; reports consistent improvements when inter-term similarity is modeled and argues that "ignored" cosine variants recover subtle semantic proximity between embedding dimensions.

By addressing both redundancy reduction and sentence prioritization, this research contributes to the development of more effective summarization techniques for educational contexts. The integration of Soft Cosine Similarity into the MMR framework represents a novel approach that has the potential to significantly improve the quality of automatic summaries, making them more useful for students and educators alike. The combination of MMR framework with SCM improves the content selection process because it selects unique valuable information that traditional generating methods often miss. This research generates practical consequences which benefit institutions more than individual learners. Learning management platforms in educational institutions can integrate these systems to enable students for automated note and study material improvement. A universal platform which enables students to enhance their academic growth will improve their study practices and increase educational success through robust educational content delivery.

Through this study focuses on using advanced technology to improve educational content through its objectives. The research combines SCM and MMR for achieving multiple essential outcomes:

- To develop an effective framework for reducing redundancy in student lecture notes.
- To enhance the prioritization of sentences to spotlight critical information.
- To implement SCM and MMR algorithms to improve the accuracy of lecture note generation.

2 METHODOLOGY

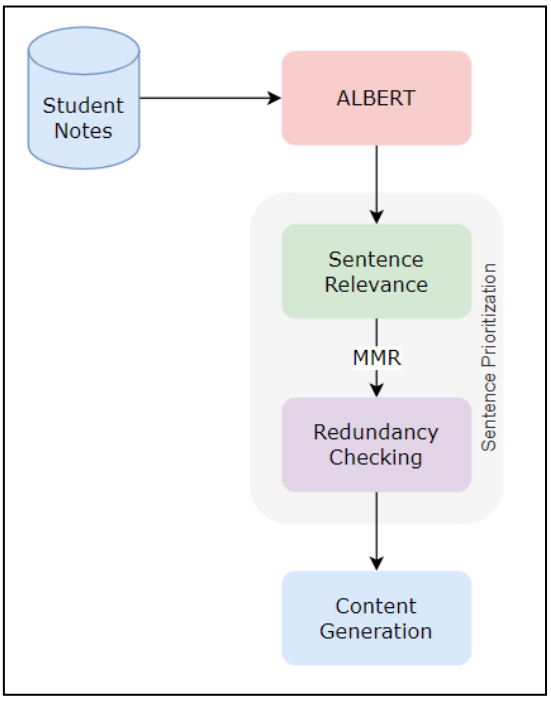
The methodologies employed to reduce redundancy and enhance the sentence relevance of student lecture notes. These methods aim to refine how educational content is processed, making it more meaningful and beneficial for students' learning processes.

2.1 System architecture

The system architecture implemented for this study is meticulously designed to handle the significant computational demands posed by the SCM and MMR algorithms, essential for the effective content generation of student lecture notes. As depicted in Figure 1, architecture integrates various components that work synergistically to process and analyze educational content.

Figure 1

SCM-MMR Workflow



2.2 Dataset

The dataset used here is the theoretical handwritten notes of the students on various subjects on different topics. In educational curriculum while the teachers are taking classes the students have the good practice of taking notes on the subject what they have understood, usually called as lecture notes. This is the best result of how far a concept is understood by each student when it is given for evaluation and this methodology is the initial step towards it. This great accomplishing move will be a new tenure as a contribution for the comprehension analysis of the students. For the purpose this study collected handwritten notes of students of grade XII on ‘English’, grade X on ‘Social Science’, BBA 'Personal Management and Human Resource Management’, MSW in social Work Practice with Individuals and Families’ and ‘Social Science Foundations for Social Work’.

2.3 Data preprocessing

Data preprocessing is a vital step in the methodology of this study, serving to prepare the collected lecture notes for effective analysis. This phase transforms raw data into a clean, standardized format suitable for computational processing, ensuring that the data analysis is both efficient and reliable.

The preprocessing of lecture notes begins with the digitization of handwritten materials. This initial step involves transcribing handwritten notes into a digital format, a critical process that not only facilitates easier manipulation and analysis but also helps in preserving the accuracy of the original student inputs. During this transcription, great care is taken to maintain the integrity of the content, ensuring that the nuances of student expressions and terminologies are not lost.

In this step, ALBERT (A Lite BERT) is employed for tokenization and embedding. Unlike traditional methods, ALBERT performs tokenization in a context-aware manner, breaking down complex texts into tokens that represent subword units. This allows the tokenization process to be sensitive to the syntactic and semantic nuances present in the text, providing tokens that are rich in contextual information. This is particularly beneficial for processing educational content where accuracy and context are paramount.

2.4 Integration of SCM and MMR

The integration of SCM and MMR into a single framework is central to this methodology. SCM is used to measure the semantic similarities between terms in the lecture notes, which helps in identifying and grouping similar content. MMR is then applied to these groups to select the most relevant sentences while minimizing redundancy. This dual approach ensures that the final generated content of lecture notes is not only concise but also comprehensive, covering all necessary topics discussed in the lectures without repetitive information.

2.4.1 Concept mapping

Concept mapping in this study involves the visualization and analysis of the relationships among concepts extracted from student lecture notes. This process is integral to understanding how students comprehend and structure information, providing insights that enhance the application of the SCM and MMR algorithms.

The mathematical framework for concept mapping can be understood through graph theory, where concepts are represented as nodes and relationships between them as edges. The graph is defined as follows:

Let $G = (V, E)$ represent the graph, where V is the set of vertices (or nodes) and E is the set of edges (or links) between them.

- Each node $v_i \in V$ represents a unique concept identified in the student notes.
- Each edge $e_{ij} \in E$ represents a relationship between concepts v_i and v_j, such as semantic similarity or contextual relevance.

The adjacency matrix, A, of graph G is used to represent which nodes are adjacent to each other (i.e., directly connected by an edge). The matrix is defined where each element a_{ij} is 1 if there is an edge between v_i and v_j, and 0 otherwise.

The process of concept mapping involves several steps:

1. *Identification of Concepts:* Extract key concepts from the lecture notes. This extraction is often supported by NLP techniques that identify significant terms or phrases within the text.
2. *Establishment of Relationships:* Determine how these concepts are related based on their occurrences and contexts within the notes. Relationships could be based

on direct mentions, semantic similarities, or co-occurrences in specific sections of the notes.

3. *Graph Construction:* Construct graph G by setting nodes for each concept and edges for each identified relationship.
4. *Visualization and Analysis:* Use graph visualization tools to depict the concept map, enabling visual analysis of how concepts are interconnected. This visualization helps in understanding the structure of knowledge as perceived and noted by students.

By systematically mapping out the relationships between concepts in lecture notes, the study leverages these insights to better apply SCM for measuring semantic similarities and MMR for evaluating the relevance and uniqueness of content.

2.4.2 Contextual tokenization

Contextual tokenization is a sophisticated method used in the processing of student lecture notes to break text into tokens that are semantically meaningful within their specific contexts. This process goes beyond basic tokenization by incorporating the understanding of the context in which words appear, which is crucial for tasks like sentiment analysis, named entity recognition, and content generation. The mathematical model for contextual tokenization can be described using probabilistic models that maximize the likelihood of a sequence of words given their surrounding context:

1. Definition of the Sequence and Context:

Let $W = \{w_1, w_2, \ldots, w_n\}$ represent a sequence of words.

The objective is to maximize the probability $P(W|\text{context})$

2. Modeling Contextual Dependencies:

Contextual dependencies are modeled using techniques like Hidden Markov Models (HMM), Conditional Random Fields (CRFs), or neural network architectures such as Recurrent Neural Networks (RNNs) or Transformers.These models consider the sequence of words and utilize the context to predict the next word in the sequence.

3. Application of Contextual Embeddings:

Modern approaches utilize deep learning models like Transformers which employ an attention mechanism to weigh the importance of each word in the context of others in the sequence.

The self-attention mechanism is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Where Q, K, V are queries, keys, and values matrices respectively, and d_k is the dimensionality of the keys.

4. Calculation of Contextual Token Probabilities:

The probability of a token sequence given the context is calculated using the context-aware embeddings generated by the self-attention mechanism:

$$P(W|\text{context}) = \prod_{i=1}^{n} P(w_i|w_1, \ldots, w_{i-1}; \text{context}) \quad (2)$$

This probability reflects how likely a word sequence is, given the preceding words and the specific contextual understanding developed by the model.

5. Optimization:

The parameters of the models are optimized to maximize $P(W|\text{context})$ across all sentences in the dataset, ensuring that the tokens generated are not only contextually relevant but also contribute significantly to the understanding of the text.

2.4.3 Vectorization

Vectorization in the context of processing student lecture notes involves converting the contextual tokens derived from the textual data into a numerical format that machine learning models can interpret and analyze. This numerical representation allows for the application of various algebraic and geometric operations on the text data, facilitating advanced computational tasks such as similarity measurement, relevance scoring, and redundancy detection, essential for effective content generation and information retrieval in educational settings.

1. Representation of Tokens as Vectors:

Let $T = \{t_1, t_2, \ldots, t_m\}$ represent the set of tokens obtained from contextual tokenization.

Each token t_i is transformed into a vector v_i in a high-dimensional vector space.

2. Vector Space Model:

The vector space is defined such that each dimension corresponds to a feature of the token, which could be a word, part of speech, or a semantic property.

The vectors are typically formed using methods such as TF-IDF or word embeddings:

$$v_i = \text{embed}(t_i) \tag{3}$$

$\text{embed}(t_i)$ is a function that retrieves or calculates the vector representation of t_i

3. Calculation of Term Frequency-Inverse Document Frequency (TF-IDF):

Term Frequency (TF) for a token t_i in a document d is calculated as:

$$\text{TF}(t_i, d) = \frac{\text{count of } t_i \text{ in } d}{\text{total number of tokens in } d} \tag{4}$$

Inverse Document Frequency (IDF) for a token t_i is calculated across a corpus D as:

$$\text{IDF}(t_i, D) = \log\left(\frac{\text{total number of documents in } D}{\text{number of documents containing } t_i}\right) \tag{5}$$

The TF-IDF score for t_i in document d is then:

$$\text{TF-IDF}(t_i, d) = \text{TF}(t_i, d) \times \text{IDF}(t_i, D) \tag{6}$$

4. Application of Word Embeddings:

Word embeddings such as Word2Vec, GloVe, or FastText can be used to convert tokens into vectors that capture semantic meanings:

$$v_i = \text{Word2Vec}(t_i) \quad \text{or} \quad v_i = \text{GloVe}(t_i) \tag{7}$$

These embeddings are pre-trained on large corpora and are capable of capturing contextual nuances in the vector space.

5. Normalization of Vectors:

To ensure consistency in magnitude, vectors are often normalized to unit length:

$$v_i' = \frac{v_i}{|v_i|} \quad (8)$$

This normalization helps in maintaining numerical stability and improves the performance of cosine similarity measures used later in SCM.

2.4.4 Sentence relevance

Sentence relevance in the context of content generation and text analysis involves determining the importance or pertinence of sentences within a document to ensure that the most significant content is identified and utilized. This process uses a combination of numerical measures and algorithms to score sentences based on their contribution to the overall document's meaning. By applying these scores in the analysis, the system ensures that the content generation or information retrieval processes focus on the most informative parts of the text. This methodical approach aids in enhancing the precision and effectiveness of the content processing tasks, particularly in educational and research settings where the extraction of pertinent information is crucial.

1. Sentence Scoring:

Let $S = \{s_1, s_2, \ldots, s_n\}$ represent the set of all sentences in a document.

Each sentence s_i is transformed into a vector s_i using the vectorization techniques described previously.

The relevance score r_i for each sentence s_i is calculated based on its semantic contribution to the document's theme.

2. Calculation of Sentence Relevance:

The relevance of a sentence can be quantified by its cosine similarity to a document vector d that represents the entire document:

$$r_i = \frac{s_i \cdot d}{|s_i||d|} \quad (9)$$

d is typically the mean of the vectors of all sentences in the document or a weighted mean where weights are derived from TF-IDF scores.

3. Normalization of Relevance Scores:

To ensure comparability across different documents or sections, relevance scores are often normalized:

$$r_i' = \frac{r_i - \min(R)}{\max(R) - \min(R)} \quad (10)$$

Here, $R = \{r_1, r_2, \ldots, r_n\}$ are the raw relevance scores, and r_i' is the normalized score of sentence s_i.

4. Weighted Relevance:

In contexts where certain sentences are deemed more critical based on their position (e.g., headings, opening sentences), a weight w_i can be applied:

$$r_i'' = w_i \times r_i' \quad (11)$$

Weights w_i can be determined by additional metadata or structural cues within the document.

5. Aggregation of Relevance Scores:

For documents with multiple sections or themes, relevance scores can be aggregated to determine the overall importance of sentences across the document:

$$R_{\text{total}} = \sum_{i=1}^{n} r_i'' \quad (12)$$

This aggregation helps in identifying sentences that are consistently relevant across different parts of the document.

2.4.5 Redundancy checking

Redundancy checking is crucial in text content generation and information retrieval to ensure that the content is concise and non-repetitive. This process involves identifying and eliminating sentences or phrases that do not add new information but rather repeat what has already been stated.

1. Vector Representation of Sentences:

Let $S = \{s_1, s_2, \ldots, s_n\}$ represent the set of sentence vectors in the document, where each sentence s_i has been vectorized as described previously.

2. Pairwise Sentence Similarity Calculation:

Calculate the cosine similarity between each pair of sentence vectors to determine the degree of redundancy:

$$\text{sim}(s_i, s_j) = \frac{s_i \cdot s_j}{|s_i||s_j|} \quad (13)$$

This measure indicates how similar two sentences are, with a higher value indicating higher redundancy.

3. Threshold for Redundancy:

Define a similarity threshold θ such that if $\text{sim}(s_i, s_j) \geq \theta$, sentences s_i and s_j are considered redundant.

4. Redundancy Elimination:

For each sentence s_i, compare its similarity with all other sentences:

$$\text{redundant}(s_i) = \bigvee_{j \neq i} (\text{sim}(s_i, s_j) \geq \theta) \quad (14)$$

If $\text{redundant}(s_i)$ is true for any j, then s_i is marked for exclusion from the final generated content.

5. Aggregation of Non-Redundant Content:

Collect all sentences for which $\text{redundant}(s_i)$ is false:

$$S_{\text{nr}} = \{s_i \in S \mid \neg \text{redundant}(s_i)\} \quad (15)$$

S_{nr} is the set of sentences that will be included in the final generated content, ensuring diversity and novelty.

By systematically applying these mathematical equations for redundancy checking, the process ensures that the final text output is devoid of unnecessary repetition, enhancing the clarity and brevity of the content. This step is particularly important in educational settings where concise and direct communication of information is crucial for effective learning and retention.

2.4.6 Maximal marginal relevance (MMR)

MMR is a technique used to balance the relevance and diversity of selected sentences in text content generation tasks, ensuring that the output is both informative and non-redundant. The content generation process effectively incorporates the most relevant content while ensuring that the information presented is varied and covers different aspects of the topic without redundancy. This method is particularly valuable in educational and research settings where comprehensive yet concise generated content are essential.

1. Definition of Relevance and Diversity:

Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of all sentence vectors in a document. Let D be the vector representing the overall document or the query in context.

2. Relevance Calculation:

The relevance of a sentence s_i to the document D is computed using cosine similarity:

$$\text{rel}(s_i, D) = \frac{s_i \cdot D}{|s_i||D|} \tag{16}$$

3. Diversity Calculation:

The diversity of a sentence s_i relative to a set of selected sentences $S' \subset S$ is calculated by determining its dissimilarity to sentences already included inS' :

$$\text{div}(s_i, S') = 1 - \max_{s_j \in S'} \left(\frac{s_i \cdot s_j}{|s_i||s_j|} \right) \tag{17}$$

4. MMR Score Calculation:

The MMR score for each sentence s_i is calculated by combining its relevance and diversity scores, controlled by a parameter λ , which selects sentences that balance relevance with the minimization of redundancy:

$$\text{MMR}(s_i) = \lambda \times \text{rel}(s_i, D) + (1 - \lambda) \times \text{div}(s_i, S') \tag{18}$$

λ is a parameter in the range $[0,1]$.

Compute the variance of ALBERT-generated sentence embeddings to gauge the semantic diversity within each document.

Dynamic λ Adjustment:

$$\lambda = \frac{1}{1+e^{-k(d-t)}} \tag{19}$$

where d is the semantic diversity, t is a pre-determined threshold, and k adjusts the response sensitivity of λ to changes in diversity.

5. Sentence Selection Process:

In each iteration of the content generation process, select the sentence s^* that maximizes the MMR score:

$$s^* = \arg \max_{s_i \in S \setminus S'} \text{MMR}\,(s_i) \tag{20}$$

Add s^* to the set S' and remove s^* from S.

6. Iterative Selection:

Continue selecting sentences based on the MMR score until a stopping criterion is met, such as a maximum content length or a minimum MMR threshold.

2.5 Algorithm

The algorithm developed for this study operates in several stages. After preprocessing, the SCM algorithm assesses the semantic similarities across the lecture notes. The output from SCM serves as the input for the MMR algorithm, which then evaluates both the relevance and uniqueness of each sentence in the context of the entire document set. The algorithm uses a scoring system to prioritize sentences that add new information and are crucial for a holistic understanding of the lecture content. The final output is a generated version of the lecture notes that maintains the educational integrity of the original content while being more digestible and time-efficient for student review.

Algorithm: Redundancy Reduction and Sentence Prioritization

Input:

- $L = \{l_1, l_2, \ldots, l_m\}$: Set of raw lecture notes

- λ: Parameter to balance relevance and diversity in MMR

Output:

- S': Set of sentences constituting the content generation

Procedure:

1. Preprocessing and ALBERT Embedding:

$$P = \emptyset$$

For each note l_i in L:

$p_i = \text{ALBERT\_Preprocess}(l_i)$ (Tokenization and normalization using ALBERT)

Add p_i to P

2. ALBERT Vectorization:

$$V = \emptyset$$

For each p_i in P:

$v_i = \text{ALBERT\_Embed}(p_i)$ (Generate contextual embeddings for p_i using ALBERT)

Add v_i to V

3. Sentence Extraction and Initial Scoring:

$$S = \emptyset$$

For each v_i in V:

$$S_i = \text{ExtractSentences}(v_i)$$

For each sentence s in S_i:

$\text{score}(s) = \text{ComputeInitialScore}(s, D)$ (Use ALBERT embeddings for scoring)

Add s to S

4. Redundancy Removal and Relevance Calculation:

$$S_{nr} = \emptyset$$

For each s in S:

If $\text{IsRedundant}(s, S_{nr}) < \theta$ (Check redundancy using contextual similarity from ALBERT embeddings):

Add s to S_{nr}

5. MMR-Based Selection:

$$S' = \emptyset$$

While $|S'| <$ MaxContentSize:

$s^* = \arg\max_{s \in S_{nr} \setminus S'} \left(\lambda \times \text{rel}(s, D) + (1\lambda) \times \text{div}(s, S')\right)$ (Select sentences using MMR with relevance and

diversity computed from ALBERT embeddings)

Add s^* to S'

Remove s^* from S_{nr}

6. Output Generation:

Return S'

Functions Used:

- $Preprocess(l)$: Normalize, tokenize, remove stopwords, and lemmatize the input lecture note l.
- $Vectorize(p)$: Convert preprocessed text p into a vector representation.
- $ExtractSentences(v)$: Segment the vectorized text v into individual sentences.
- $ComputeInitialScore(s, D)$: Calculate the initial relevance score of sentence s based on document D.
- $IsRedundant(s, S_{nr})$: Check if sentence s is redundant with respect to sentences in S_{nr}, returning a similarity score.
- $rel(s, D)$: Compute the relevance of sentence s relative to the overall document D.
- $div(s, S')$: Compute the diversity of sentence s relative to selected set S'.

3 RESULTS AND DISCUSSION

3.1 Parameter settings

Table 1 represents the parameter settings for the proposed SCM-MMR model.

Table 1

Parameter Dimensions and Embeddings

Parameter	Description	Value/Setting
Embedding Dimensions	The size of the vector space for ALBERT embeddings.	768 dimensions (typical for ALBERT-base)
Similarity Threshold	Minimum cosine similarity for terms to influence SCM.	0.7
Lambda (λ)	Balances relevance and diversity in MMR.	0.65

Relevance Function	Measures how relevant a sentence is to the document context.	Cosine similarity to the document vector
Diversity Function	Measures how different a sentence is from those selected.	Minimum cosine similarity to selected sentences

3.2 Results of collected datasets

Table 2 presents the results of the proposed SCM-MMR model, showcasing its effectiveness in generating educational content across various academic levels and subjects as evidenced by the ROUGE metrics: ROUGE-W, ROUGE-L, and ROUGE-N. These metrics provide a comprehensive evaluation by focusing on word overlap (W), longest common subsequence (L), and n-gram overlap (N).

The results indicate that the SCM-MMR model performs exceptionally well across all datasets, with particularly strong results in the Grade X Social Science dataset, where the model achieves ROUGE scores of 0.864 (W), 0.876 (L), and 0.901 (N). This suggests a high degree of accuracy and relevance in the content generated, capturing both the breadth and depth of the original content effectively.

Similarly, in the Grade XII English dataset, the model records ROUGE scores of 0.821 (W), 0.844 (L), and 0.863 (N), demonstrating its robust capability to handle complex linguistic structures and academic vocabulary effectively. The BBA courses, both in Personal Management (PM) and Human Resource Management (HRM), also show commendable performance with ROUGE-W scores of 0.798 and 0.851 respectively, and even higher in their L and N metrics, emphasizing the model's consistency in maintaining key informational elements.

For the more specialized MSW (Master of Social Work) courses, the generated content maintains good quality with the MSW SSFSW (Social Science Foundations for Social Work) dataset showing scores of 0.783 (W), 0.801 (L), and 0.837 (N), and the MSW SWPIF (Social Work Practice with Individuals and Families) dataset showing 0.846 (W), 0.870 (L), and 0.894 (N). These results reflect the model's adaptability to diverse content types and its effectiveness in preserving essential content generation.

Table 2

Result of the SCM-MMR model

MMR Algorithm Results	ROUGE-W	ROUGE-L	ROUGE-N
Grade XII English	0.821	0.844	0.863
Grade X Social Science	0.864	0.876	0.901
BBA PM	0.798	0.811	0.842
BBA HRM	0.851	0.865	0.882
MSW SSFSW	0.783	0.801	0.837
MSW SWPIF	0.846	0.870	0.894

Figure 2 showcases the ROUGE-N scores from the evaluation of the SCM-MMR model across various educational datasets. In the figure, the SCM-MMR model achieves the highest ROUGE-N score in the Grade X Social Science dataset, with a score of 0.901, indicating exceptional quality that likely captures essential content with great accuracy. This is followed by the Grade XII English and BBA HRM courses, where the scores are 0.863 and 0.882, respectively, suggesting that the model performs robustly across varied subject matter, including both humanities and business courses.

The results for the BBA PM and MSW courses also demonstrate the model's effectiveness, albeit with slightly lower scores. For BBA PM, the ROUGE-N score stands at 0.842, while for MSW SWPIF it is 0.894, and MSW SSFSW has the lowest score at 0.837. These results indicate that while the model is highly effective, there are variations in performance likely due to the differing complexities and terminologies inherent in each field of study.

Figure 2

Result of Rouge N of SCM-MMR Model

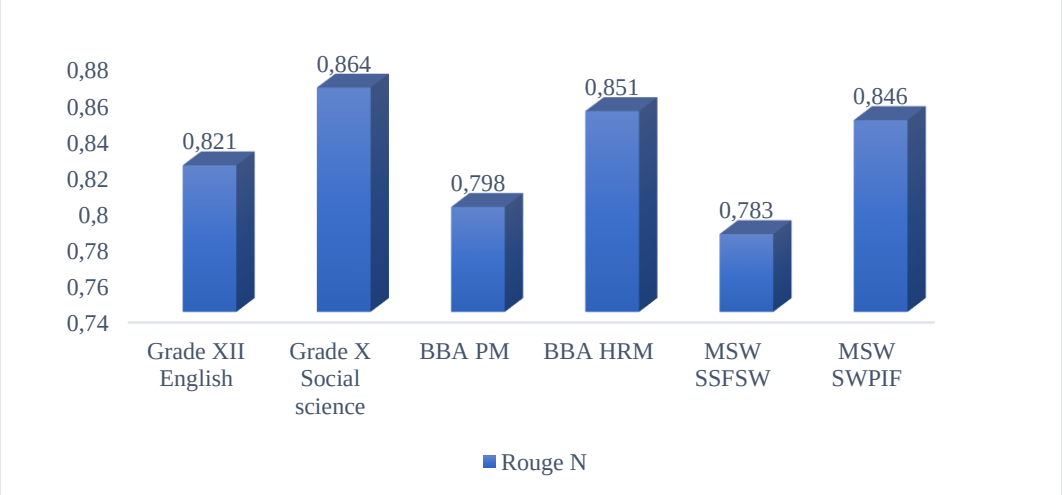
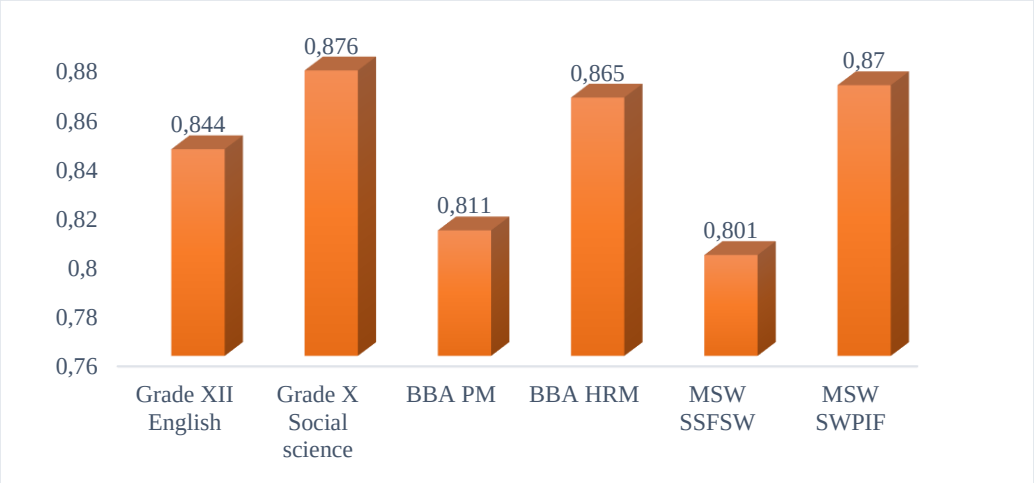


Figure 3 illustrates the ROUGE-L scores from the evaluation of the SCM-MMR model across various educational datasets. The graph illustrates the model's ability to produce academic texts by analyzing the longest common subsequence, an important measure for evaluating the consistency and coherence of the content generated.

The SCM-MMR model achieves its highest ROUGE-L score in the Grade X Social Science dataset with a score of 0.876. This exceptional performance indicates the model's strong capability in maintaining the structural integrity and essential content of the social science materials, which often involve complex discussions and thematic depth. Following closely, the MSW SWPIF (Social Work Practice with Individuals and Families) dataset shows a ROUGE-L score of 0.870.

Figure 3

Result of Rouge L of SCM-MMR Model



The BBA HRM (Human Resource Management) datas*et al*so sees a robust performance with a ROUGE-L score of 0.865, highlighting the model's proficiency in generating business and management texts that require a nuanced understanding of concepts and their applications.

The scores for Grade XII English and BBA PM (Personal Management) are 0.844 and 0.811, respectively, indicating proficient content generation but suggesting potential areas for further refinement, especially in handling the literary analysis and management strategies where specific jargon and concepts might challenge the content generation process.

The lowest score is observed in the MSW SSFSW (Social Science Foundations for Social Work) dataset at 0.801. Although this score is the lowest among the datasets presented, it still reflects a reasonable effectiveness in generating academic texts in social work, an area often characterized by its narrative depth and ethical considerations.

Figure 4 provides a visual representation of the ROUGE-N scores achieved by the SCM-MMR model across various academic courses. ROUGE-N, which measures the overlap of n-grams between the generated contents and the reference texts.

Figure 4

Result of Rouge N of SCM-MMR Model

The SCM-MMR model shows excellent performance across all datasets, with standout scores in the Grade X Social Science and MSW SWPIF (Social Work Practice with Individuals and Families) courses, achieving scores of 0.901 and 0.894, respectively. These high scores indicate that the model is exceptionally adept at capturing detailed n-gram overlaps in these disciplines, which often involve complex and nuanced content that requires accurate representation in content generation.

In the Grade XII English dataset, the model scores 0.863, demonstrating its strong capability in accurately generating literary and critical analyses, where capturing specific phrases and terminology is crucial. The BBA HRM (Human Resource Management) datas*et al*so sees a high score of 0.882, reflecting the model's effectiveness in generating business management content, which often includes specialized vocabulary and concepts.

The BBA PM (Personal Management) and MSW SSFSW (Social Science Foundations for Social Work) datasets show slightly lower but still robust scores of 0.842

and 0.837, respectively. These results suggest that while the model performs well, there may be room for improvement in capturing some of the more intricate or less frequently occurring n-grams within these specific fields.

3.3 Results of publicly available datasets

The results from Table 3 and Figure 5 offer a detailed comparison between the proposed SCM-MMR model and the RL-MMR model utilized by Mao *et al*. (2020) across TAC2011 datasets. The SCM-MMR model outperforms the Reinforcement Learning MMR (RL-MMR) model across all ROUGE metrics, which is indicative of its superior content generation capabilities. The SCM-MMR achieved ROUGE-W, ROUGE-L, and ROUGE-N scores of 0.821, 0.844, and 0.863, respectively. These results significantly surpass the RL-MMR's scores of 0.114, 0.150, and 0.396 in the TAC2011 dataset, as reported in their study.

The notable performance improvement highlighted in these results can be attributed to the integration of the SCM-MMR model. Unlike the RL-MMR, which primarily leverages reinforcement learning techniques to guide content generation, the SCM-MMR model incorporates semantic understanding and diversity management more robustly. SCM's ability to preserve semantic similarities between terms enhances the model's capability to detect and prioritize key information more effectively. Concurrently, MMR's function in optimizing the diversity of the information extracted minimizes redundancy, ensuring that the generated content are not only accurate but also concise.

These advantages are particularly beneficial in educational contexts where the clarity and quality of generated content directly impact learning outcomes. The high performance of SCM-MMR in generating complex educational materials into digestible formats suggests that it could significantly aid students and educators by providing clearer, more focused study materials.

Table 3

Comparison of the proposed model with the RL-MMR

Dataset	Model	Rouge W	Rouge L	Rouge N
TAC2011	RL-MMR	0. 114	0.15	0.396
	SCM-MMR	0.647	0.688	0.691

Figure 5

Comparison of the proposed model with the RL-MMR

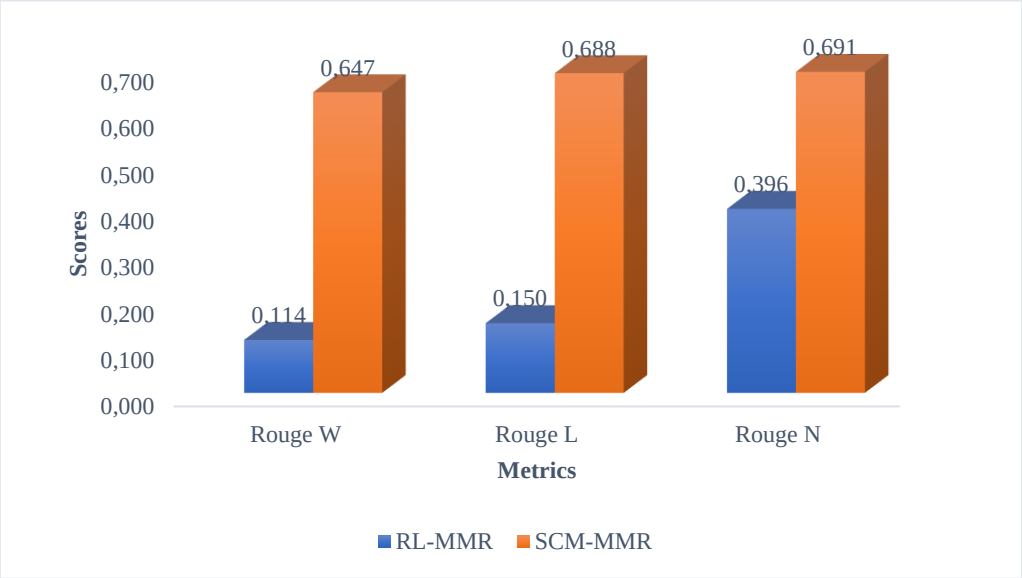


Table 4 and Figure 6 present a focused comparison of the proposed SCM-MMR model against the keyword-based content generation model discussed by He *et al*. (2020). The evaluation, conducted using the ROUGE metrics across academic datasets from arXiv scientific papers, reveals noteworthy insights into the content generation performance of the two models.

The SCM-MMR model demonstrates substantial improvements over the Controllable Text Summarization (CTRLSum) model in all evaluated metrics. Specifically, the ROUGE-W, ROUGE-L, and ROUGE-N scores of the SCM-MMR are significantly higher, showcasing its ability to produce content generation that are not only relevant but also more informative and concise. For example, in the arXiv scientific papers dataset, SCM-MMR achieves a ROUGE-W score of 0.763, a ROUGE-L score of 0.791, and a ROUGE-N score of 0.812. In contrast, the CTRLSum model scores 0.183, 0.427, and 0.475 respectively in the same metrics.

This performance disparity underscores the advanced capabilities of the SCM-MMR model. SCM enhances the semantic understanding of texts by measuring similarities beyond mere keyword overlaps, which is critical in complex domains like scientific literature and patent documents. This feature allows for a deeper comprehension of context and content relevancy, which is crucial for generating specialized academic materials accurately.

Table 4

Comparison of the proposed model with the CTRLSum model

Dataset	Model	Rouge W	Rouge L	Rouge N
arXiv scientific papers	CTRLSum	0.183	0.427	0.475
	SCM-MMR	0.763	0.791	0.812

Figure 6

Comparison of the proposed model with the CTRLSum model

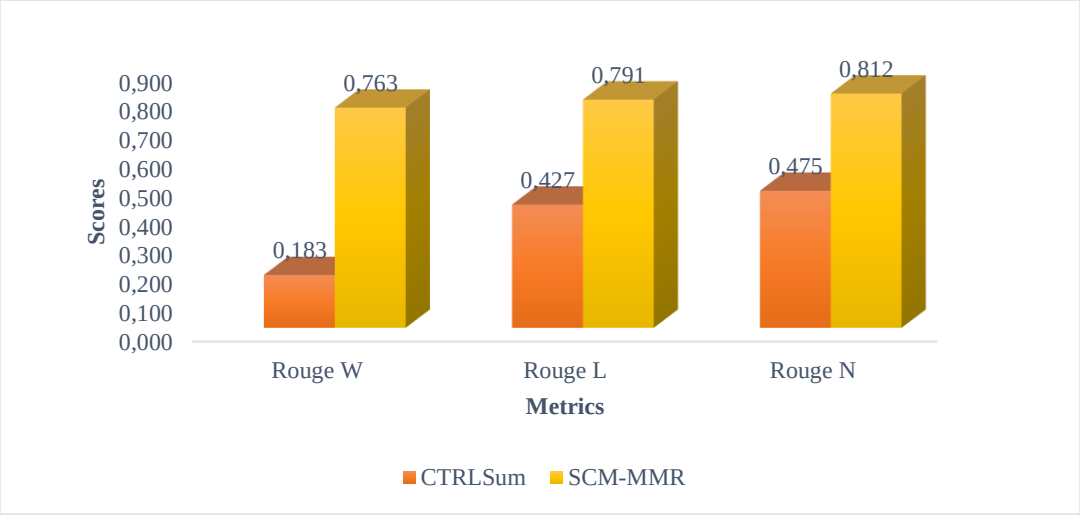


Table 5 and Figure 7 offer a comparative analysis of the proposed SCM-MMR model against the Abstractive GAN model discussed by Yang *et al*. (2021) in their study on hierarchical deep neural networks for text generation. The evaluation, conducted using ROUGE metrics across comprehensive datasets like the Gigaword corpus and CNN/daily mail corpus, highlights the effectiveness of different content generation strategies employed by these models.

In comparison, the SCM-MMR model consistently outperforms the Abstractive GAN model in all ROUGE metrics, illustrating the robustness and effectiveness of the

SCM-MMR approach in producing high-quality contents. Specifically, the SCM-MMR achieves ROUGE-W, ROUGE-L, and ROUGE-N scores of 0.602, 0.613, and 0.634 respectively, against the Abstractive GAN scores of 0.203, 0.391, and 0.431 in the same metrics. This marked improvement underscores the advantages of integrating semantic understanding and relevance assessment directly into the content generation process.

The SCM component enhances the model's ability to grasp semantic nuances by evaluating similarities based on contextual meanings rather than mere word overlap. This is particularly beneficial in handling diverse and complex datasets like news articles and press releases, where understanding the underlying context and the interrelation of concepts is crucial for generating coherent and informative content.

Furthermore, the MMR component of the SCM-MMR model optimizes the selection of sentences by balancing relevance and diversity, ensuring that the content is not only comprehensive but also free of redundant information. This contrasts with the Abstractive GAN approach, which, while innovative in generating narrative-like text, may struggle with maintaining the factual accuracy and pertinence of content as evidenced by the lower ROUGE scores.

Table 5

Comparison of the proposed model with the Abstractive GAN

Dataset	Model	Rouge W	Rouge L	Rouge N
Gigaword corpus, CNN/ daily mail corpus	Abstractive (GAN)	0. 203	0.391	0.431
	SCM-MMR	0.602	0.613	0.634

Figure 7

Comparison of the proposed model with the Abstractive GAN

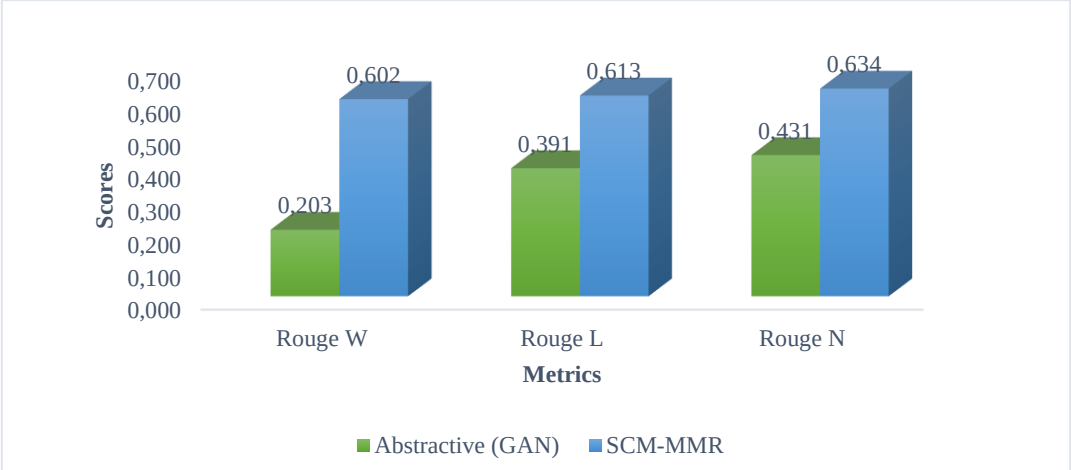


Table 6 and Figure 8 present the comparative analysis of the proposed SCM-MMR model against the Parsing Summary Fine-Tuned (ParsingSum-FT) model as studied by Ma *et al*., (2024). This evaluation, which utilizes the ROUGE metrics across the Multi-News dataset, shows the effectiveness of both models in handling complex multi-document content generation tasks.

The SCM-MMR model shows very competitive results, closely aligning with or slightly outperforming the ParsingSum-FT model on various metrics. Specifically, the SCM-MMR achieves ROUGE-W, ROUGE-L, and ROUGE-N scores of 0.668, 0.693, and 0.717 respectively, compared to the ParsingSum-FT scores of 0.657, 0.678, and 0.692. These results are indicative of the high efficacy of the SCM-MMR model in producing coherent and comprehensive content from multiple news articles.

The SCM component is particularly effective in this context as it enhances the model's ability to identify and leverage semantic similarities across different documents, which is critical in multi-document dataset were understanding the relationships between various pieces of information is key. This results in generating content that is brief yet full of essential details, preserving the key points of the news articles.

Moreover, the MMR component ensures that the final generated content is not just a collection of facts but a well-rounded narrative that emphasizes the most relevant information while minimizing redundancy. This is crucial in multi-news dataset, where the challenge is to merge information from various sources into a single, unified document that maintains the narrative flow and integrity.

Table 6

Comparison of the proposed model with the ParsingSum-FT

Dataset	Model	Rouge W	Rouge L	Rouge N
Multi-News	ParsingSum-FT	0.657	0.678	0.692
	SCM-MMR	0.668	0.693	0.717

Figure 8

Comparison of the proposed model with the ParsingSum-FT

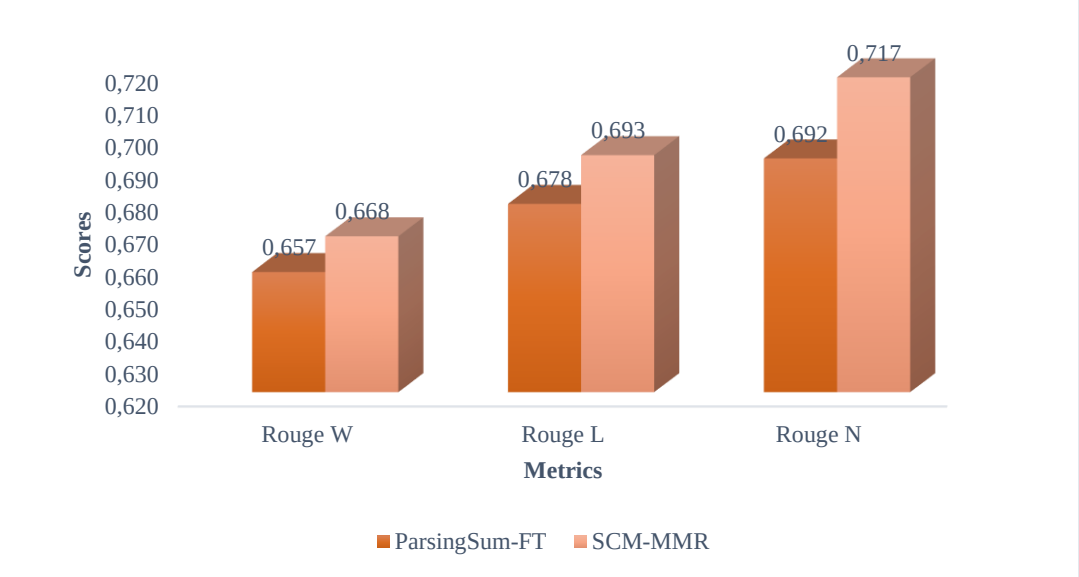


Table 7 and Figure 9 present the results of the comparison between the proposed SCM-MMR model and the mTLDRgen model developed by Atri *et al*. (2023). This evaluation uses the ROUGE metrics across the How2 dataset, which is known for its complexity due to the extensive range of topics and detailed content it includes.

In comparison, the SCM-MMR model performs remarkably well, showcasing close competition with the mTLDRgen model. Specifically, the SCM-MMR achieves ROUGE-W, ROUGE-L, and ROUGE-N scores of 0.752, 0.795, and 0.826, respectively, versus the mTLDRgen scores of 0.758, 0.789, and 0.801 in the same metrics. These scores indicate that while the mTLDRgen model slightly outperforms SCM-MMR in terms of ROUGE-W, the SCM-MMR model surpasses mTLDRgen in ROUGE-L and ROUGE-N scores.

The SCM component effectively captures the semantic nuances within the text, enhancing the model's ability to understand and process complex and varied information from the How2 dataset. This capability is crucial for maintaining the integrity and depth of content.

Moreover, the MMR component strategically selects the most relevant and diverse sentences for inclusion of content generation, ensuring that the essential information is covered without redundancy. This approach is beneficial in contexts like the How2

dataset, where there is a significant amount of data, and maintaining content diversity without losing relevance is challenging.

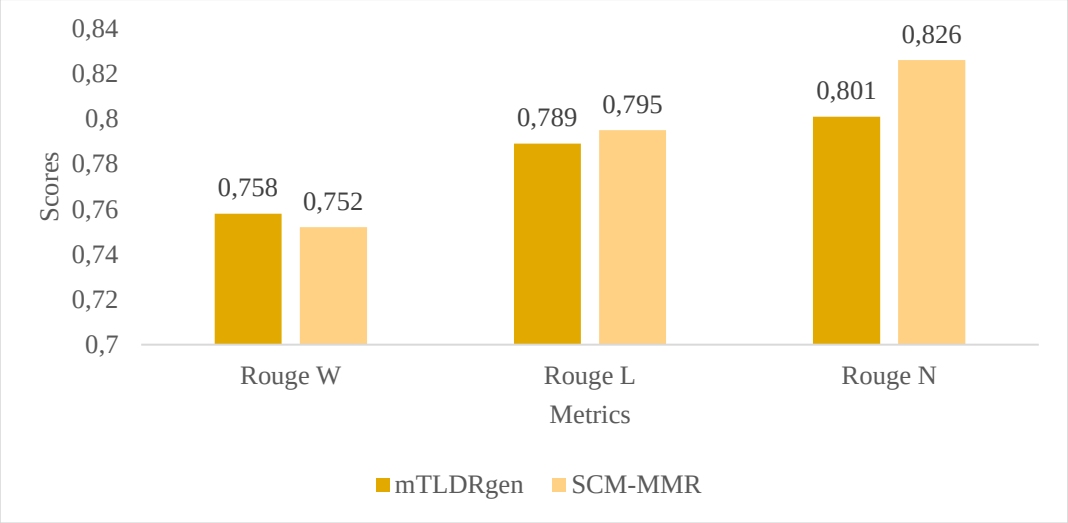
Table 7

Comparison of the proposed model with the mTLDRgen

Dataset	Model	Rouge W	Rouge L	Rouge N
How2	mTLDRgen	0.758	0.789	0.801
	SCM-MMR	0.752	0.795	0.826

Figure 9

Comparison of the proposed model with the mTLDRgen



3.4 Discussion

The SCM-MMR method enhances academic text handling, showing improvement across subjects and levels. This discussion explores its strategic implications, adaptations, and broader impacts on education and learning environments.

The SCM-MMR model, as demonstrated by the results, excels in capturing essential information from complex lecture notes, thereby streamlining the study process for students. One of the remarkable outcomes is the consistent performance across different subjects, showcasing the model's versatility and robustness. High ROUGE scores show the model captures relevant content while preserving the original text's structure, aiding student comprehension and retention.

The application of the SCM-MMR model has demonstrated a reduction in redundant information, addressing a common issue where students may be overwhelmed by extensive notes. By prioritizing and condensing key information, the model helps to enhance focus and reduce cognitive load, which can improve learning outcomes. This characteristic of the model meets current educational needs, emphasizing brevity and retention of essential information.

As educational content becomes increasingly diverse and voluminous, tools that adapt to various disciplines and content complexities are essential. The SCM-MMR model's ability to handle a wide range of subjects from social sciences to technical courses in management and human resources exemplifies its potential as a scalable solution for educational institutions.

However, while the model demonstrates substantial benefits, it also prompts considerations for future enhancements. For instance, while the model performs exceptionally well in subjects with structured content, such as social sciences, there are opportunities to fine-tune its application in more dynamically structured disciplines such as literature or complex theoretical subjects. Future iterations could explore deeper semantic analysis techniques or advanced neural network models that could further refine the content generation accuracy and relevance.

The ability of the SCM-MMR model to be integrated into educational platforms or learning management systems could transform how students interact with lecture content. By providing generated, concise versions of lectures, educational platforms can offer more digestible and accessible learning materials, especially beneficial in remote learning environments where student engagement varies widely.

4 CONCLUSIONS

This research has explored the development and application of the SCM-MMR model for generating educational content, demonstrating its efficacy through extensive evaluations across various academic datasets. The model's robust performance, highlighted by high ROUGE scores, underscores its ability to distil complex lecture notes into concise by maintaining the integrity and richness of the original material. This capability is particularly beneficial in educational settings where comprehensibility and retention of critical information are paramount.

The successful implementation of SCM and MMR technologies in the SCM-MMR framework has provided valuable insights into the potential of content generation techniques in education. The model's adaptability across different subjects from the sciences to the humanities illustrates its versatility and potential for broader application in academic environments.

REFERENCES

Alizadeh, M., & Seilsepour, A. (2025). A novel self-supervised sentiment classification approach using semantic labeling based on contextual embeddings. *Multimedia Tools and Applications* 84, 10195–10220. https://doi.org/10.1007/s11042-024-19086-y

Aranzamendez, S.G, *et al*., (2024). An Enhanced Content-based Filtering Using Maximal Marginal Relevance. *International Journal of Computing Sciences Research.* Vol. 8, pp. 3070-3087. https://doi.org/10.25147/ijcsr.2017.001.1.204

Carbonell, J., & Goldstein, J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 335–336.

Chistikov, P., & Khomitsevich, O. (2013). Improving prosodic break detection in a Russian TTS system. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8113 LNAI*(3), 181–188. https://doi.org/10.1007/978-3-319-01931-4_24

Colombo, M. (2024). Semantic Similarity Measures. In *Phenotropic Interaction* (pp. 49–69). https://doi.org/10.1007/978-3-031-42819-7_4

Demilie, W. B. (2022). Comparative Analysis of Automated Text Summarization Techniques: The Case of Ethiopian Languages. *Wireless Communications and Mobile Computing*, 1–28. https://doi.org/10.1155/2022/3282127

Erkan, G., & Radev, D. R. (2011). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal Of Artificial Intelligence Research*, *22*(1), 457–479. https://doi.org/10.1613/jair.1523

Faisal Rahutomo, Teruaki Kitasuka, & Masayoshi Aritsugi. (2012). Semantic Cosine Similarity. *The 7th International Student Conference on Advanced Science and Technology ICAST*, *4*(1), 4–5.

Gunawan, G., Fitria, F., Setiawan, E. I., & Fujisawa, K. (2023). Maximum Marginal Relevance and Vector Space Model for Summarizing Students' Final Project Abstracts. *Knowledge Engineering and Data Science*, *6*(1), 57. https://doi.org/10.17977/um018v6i12023p57-68

Ijebu, F.F., Liu, Y., Sun, C., & Usip, P.U. (2025). Soft cosine and extended cosine adaptation for pre-trained language model semantic vector analysis. *Applied Soft Computing* 169: 112551. https://doi.org/10.1016/j.asoc.2024.112551

Jain, M., & Rastogi, H. (2020). Automatic Text Summarization using Soft-Cosine Similarity and Centrality Measures. *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, 1021–1028. https://doi.org/10.1109/ICECA49313.2020.9297583

Januzaj, Y., & Luma, A. (2022). Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words. *International Journal of Emerging Technologies in Learning*, *17*(12), 258–268. https://doi.org/10.3991/ijet.v17i12.30375

Jiang, P., & Cai, X. (2024). A Survey of Text-Matching Techniques. *Information* 15(6): 332. https://doi.org/10.3390/info15060332.

Kaur, N., (2024). A Review on String-Based Text Similarity Techniques in Computational Analysis. *International Journal of Intelligent Systems and Applications In Engineering* 12(23s), 3138–3144.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2011). *Mining of Massive Datasets*. Cambridge University Press. https://doi.org/10.1017/CBO9781139058452

Locke, E. A. (2015). An empirical study of lecture note taking among college students. *Journal of Educational Research*, *71*(2), 93–99. https://doi.org/10.1080/00220671.1977.10885044

Mao, Y., Qu, Y., Xie, Y., Ren, X., & Han, J. (2020). Multi-document summarization with maximal marginal relevance-guided reinforcement learning. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1737–1751. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.136

Prasetya Wibawa, A., & Kurniawan, F. (2024). A survey of text summarization: Techniques, evaluation and challenges. *Natural Language Processing Journal*, *7*, 1–21. https://doi.org/10.1016/j.nlp.2024.100070

Renkl, A., & Atkinson, R. K. (2003). Structuring the Transition From Example Study to Problem Solving in Cognitive Skill Acquisition: A Cognitive Load Perspective. *Educational Psychologist*, *38*(1), 15–22.

Sc, I. M., Science, C., Intelligence, A., Science, C., & Science, D. (n.d.). *Integrated M . Sc . Programmes in Computer Science*.

Sidorov1, G., Gelbukh1, A., Gomez-Adorno1, H., & Pinto2, D. (2014). Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Computacion y Sistemas*, *18*(3), 491–504. https://doi.org/10.13053/CyS-18-3-2043

Upadhay, N., & Singh, U. (2020). A Review on Requirements Prioritization Techniques. *International Journal of Creative Research Thoughts*, *8*(12), 877–881. https://www.researchgate.net/publication/358962528_A_Review_on_Requirements_Prioritization_Techniques

Wang, Z., Zhang, H., Chen, J., & Chen, H. (2024). An effective framework for measuring the novelty of scientific articles through integrated topic modeling and cloud model." *Journal of Informetrics* 18(4): 101587. https://doi.org/10.1016/j.joi.2024.101587

Authors' Contribution

Both authors contributed equally to the development of this article.

Data availability

All datasets relevant to this study's findings are fully available within the article.

How to cite this article (APA)

Baby, A., V, V., & Jose, J. REDUNDANCY REDUCTION AND SENTENCE PRIORITISATION OF THE STUDENT LECTURE NOTES USING SOFT COSINE IMPLEMENTED MMR ALGORITHM. *Veredas Do Direito*, e223612. https://doi.org/10.18623/rvd.v22.n4.3612