ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2025.1834.918

CONTEXT-AWARE RULE-BASED MATH EXPRESSION NORMALISER AND VERBALIZER USING LATEX2TEXT FOR ENHANCED DOCUMENT PREPROCESSING

J. Joice[1*], Dr.C. Sathya[2]

[1]Research Scholar, PG and Research Department of Computer Science, Government Arts and Science College, Kangeyam, Tiruppur, Tamil Nadu, India.*
e-mail: joymca93@gmail.com, orcid: https://orcid.org/0009-0004-2339-1881
[2]Assistant Professor, PG and Research Department of Computer Science, Government Arts and Science College, Tiruppur, Tamil Nadu, India.
e-mail: sathyavenkateswaran@gmail.com, orcid: https://orcid.org/0009-0008-1555-2664

Received: September 25, 2025; Revised: November 01, 2025; Accepted: December 05, 2025; Published: December 30, 2025

SUMMARY

Blind students usually are subjected to a substantial impediment of reading and accessing electronic documents, especially data that are noisy and those that are carefully designed. Traditional NLP models severely underestimate or misinterpret mathematical expressions in which symbols are represented as notation. It is a critical problem in the educational field, accessibility, and report generation programs, where in-depth knowledge of mathematical content is a priority. State-of-the-art document summarisation systems tend to fail in noisy text, disordered document structures, and non-textual content, e.g., equations, images, and charts. This paper introduces a powerful preprocessing model that focuses on improving input quality, semantic coherence, and readability. The process consists of sophisticated text cleaning, discerning structuring, and an extensive content interpretation model. The paper presents a proposal to simplify and verbalise mathematical expressions using a rule-based, context-sensitive language called the Verbalizer Rule (VR). The system translates complex mathematical syntax into human-readable natural-language descriptions by pattern-matching expressions and translating semantic meaning using clues in the context. Experiments demonstrate that this method achieves much higher readability scores and summarisation quality than state-of-the-art models. In the assessment, the Proposed CARMEN model, using the ROUGE metrics 1, 2, and L, yields a ROUGE score of more than 0.8333 among the other verbalizers.

Key words: *pre-processing formulae, symbolic math, LaTeX2Text, textual representation, NLP.*

INTRODUCTION

The digitisation of educational material has made it more accessible to a broader range of people; however, the visually impaired students and scholars continue to experience difficulties with accessing unstructured and complex digital materials. Students encounter difficulties in retrieving academic materials, especially those that rely heavily on visual presentation, owing to Visual impairment. We used to assist them with this summary. Document summarisation provides a significant solution to them, yet its quality and the form of the input data influence its performance. The instructional resources used in online learning of the Science, Technology, Engineering, and Mathematics (STEM) disciplines often present complex subjects through diagrams, equations, and geometric shapes that blind and sightless students cannot comprehend. Formulas cannot be vocalised correctly, as OCR algorithms, such as translation tools, cannot comprehend their meaning structure. Although there has been improvement in the development of adaptive technologies to accommodate visually impaired students, more accessible and affordable technologies should be developed to enable all students to access academic materials equally, regardless of their field of study [32]. Automatic summary generation research has advanced significantly, with both fixed and dynamic methods leveraging computer vision and artificial intelligence algorithms [1][20]. Nonetheless, there are still problems, including the prohibitive cost of refreshable tactile displays and the loudness of data creation from non-textual media. The need for a more sophisticated solution is supported by the fact that familiar TTS readers, including Adobe Acrobat and Microsoft Edge, do not recognise or read mathematical formulae correctly. Other tools attempt to convert LaTeX code to spoken English by hand, which are generally not helpful and struggle to handle exceptional cases. The study proposes a preprocessing paradigm that improves text input through multi-level processing. Academic articles use mathematical equations as a vital element, especially in engineering, physics, and computer science [27]. Nevertheless, there are a few challenges and considerations that accompany working with equations when working on an academic literature review:

- Mathematical expression of equations is not usually consistent across different articles, and that makes comparing and synthesising a result more difficult [2]. Some papers are in LaTeX format, whereas others use images as equations or plain text, making extraction and analysis tedious [3].

- The application of equations can be an individual interpretation, and they may need subject knowledge to distinguish between their applicability and use [6]. This is especially true of transdisciplinary research, where the equations can take on different meanings and interpretations across disciplines.

- Basic concepts of multivariate statistical analysis and mathematical modelling are still used on many fronts, such as supply chains and logistics [4]. The description of the algorithms to be analysed and the visualisation of the equations, however, should be enhanced to be readable and reproducible.

While equations are prominent in the depiction of high-level ideas in research papers, their presentation in literature reviews requires careful thought. Researchers will aim to be consistent in depictions of equations, include distinct descriptions of the mathematical concepts, and be clear when reporting methodology [7][5]. It would be helpful for future research to develop standardised protocols for the extraction and analysis of equations in literature reviews, possibly using NLP and Machine Learning (ML) methods that are increasingly effective and accurate [21][28].

RELATED WORK

Previous studies have explored OCR correction, language normalisation, and Equation interpretation for accessibility. [8] note that the majority of known techniques for algorithmic solution of math word problems utilise frame-based models. To mitigate shortcomings in existing models, the authors propose a new model that incorporates extended semantic networks to capture the mathematical structure in word problems. Their Solver for Mathematical Text Problems (SoMaTePs) recognises math problems using NLP, translates them into mathematical equations, and solves them using a computer algebra program. [9] mention a shift from template-based approaches with pre-specified rules to more sophisticated neural network models for generating math word problems. The authors propose a new neural network model that incorporates equation and topic information through a fusion mechanism, along with an entity-enforced loss to ensure relevance between the generated problem and equation. Previous studies focused on template- and frame-based models; more recent studies have evolved towards more sophisticated approaches using semantic networks, NLP, and neural networks to interpret and generate mathematical equations. The shift follows the rising complexity and capabilities of mathematical interpretation systems. Most systems, however, solve these tasks separately and without overall preprocessing unique to document summarisation. Multimodal assistive systems and parsing of structured documents are examples that impact our integrated approach.

RESEARCH METHODOLOGY

Data preprocessing is a vital step in preparing online education documents, especially those containing equations, for efficient machine learning and natural language processing[19][22]. Preprocessing these documents involves specialised techniques for handling mathematical expressions without loss of semantic meaning [16]. Preprocessing equations in online education documents is a difficult task. Unlike regular text, equations involve special characters, symbols, and structures that require careful handling and attention to detail. Regular text preprocessing techniques, such as tokenisation, stop word removal, and stemming, may not be applied directly to equations [10], as shown in Figure 1.

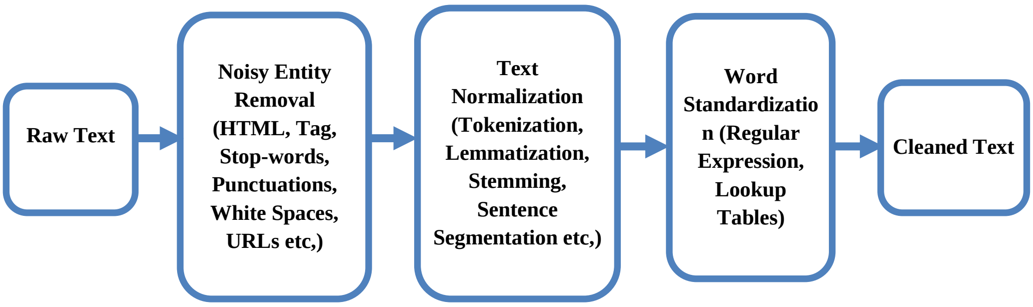

Raw Text
Noisy Entity Removal (HTML, Tag, Stop-words, Punctuations, White Spaces, URLs etc,)
Text Normalization (Tokenization, Lemmatization, Stemming, Sentence Segmentation etc,)
Word Standardizatio n (Regular Expression, Lookup Tables)
Cleaned Text

Figure 1. Text cleaning steps

Instead, specialised techniques such as mathematical expression identification and preservation, equation conversion to standard form, and handling of LaTeX or MathML notations may be necessary [11]. General text preprocessing techniques are a foundation, but translating these techniques into equation-rich online education documents is needed [12]. The creation of custom preprocessing pipelines capable of processing both textual content and mathematical expressions is key to improving the performance of summarisation models in education. The focus of research should be on developing specialised preprocessing techniques for equation-rich documents, possibly incorporating mechanisms to automatically identify and process mathematical content.

Proposed Model

Recommended Context-Aware Rule-based Mathematical Expression Normaliser (CARMEN) is a rule-based preprocessor that is built to identify, simplify, and texturise mathematical expressions in scientific or scholarly papers.

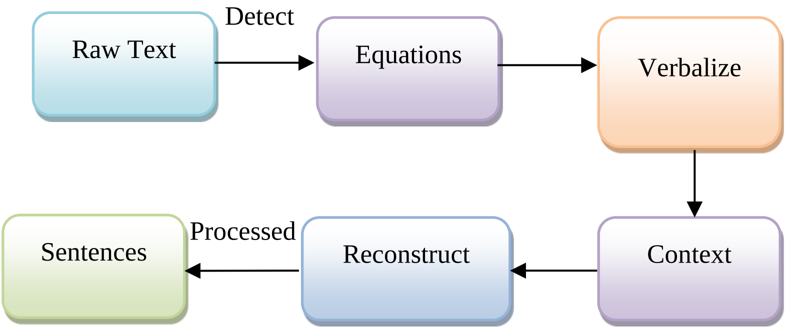


Figure 2. Proposed model flowchart

The proposed model processes the raw text into cleaned expressions, which are then passed through rule-based verbalisation, context-sensitive enrichment, and reintegration, as shown in Figure 2. It employs rule-based reasoning. It is context-sensitive, producing natural-language approximations of symbolic mathematics. It improves readability and aligns with NLP applications, such as summarisation. Its main task is to detect and extract exact mathematical expressions from scientific and scholarly papers [33]. The proposed model detects mathematical expressions in various documents using regular expressions (regex) and syntactic parsing. For the regex pattern, used to find the expression matching any text enclosed between two dollar signs $, which is the standard LaTeX syntax for inline math. The rules set are intended for mapping math style as outlined in Table 1.

Table 1. Regex-based pattern matching

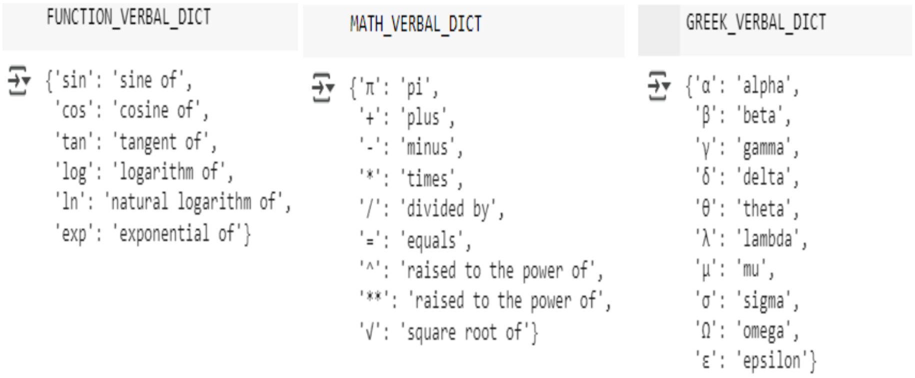
Expression Type	**Pattern**	**Regex Match**
Inline Math	\$.*?\$	$a^2 + b^2 = c^2$
Block Math	\\\ [. *? \\\] or \\begin{equation}. *? \\end{equation}	\ [\int_0^1 x^2 dx\]
Math Functions	\w+\ (. *? \)	sin(x), f(x)

The Mathematical expression applies rule-based simplification to transform expressions into more readable and word-friendly forms [13] (e.g., fractions, exponents, or nested functions into flattened expressions). Table 2 shows the symbolic expressions with readable contents.

Table 2. Expression to verbalisation

Expression Type	**Pattern (LaTeX)**	**Verbalization**
Fraction	\Frac{a}{b}	"a divided by b"
Power/Exponentiation	x^n	"x raised to the power n"
Roots	\sqrt{a}	"square root of a"
Integrals	\int_a^b f(x) dx	"the integral off of x from a to b"
Summation	\sum_{i=1} ^n a_i	"the sum of a sub i from i equals 1 to n"
Limits	\lim_{x \to 0} f(x)	"the limit off of x as x approaches zero"
Derivatives	\frac{dy}{dx}	"the derivative of y with respect to x"

This also accommodates algebraic normalisation rules. Converts the simplified expressions to natural language strings (e.g., x^2 + y^2 = r^2 to "x squared plus y squared equals r squared") using a mapping table and grammar-sensitive verbalisation patterns. Dictionaries are constructed to convert mathematical symbols, functions, and patterns into spoken/written equivalents. Dictionaries include basic symbols (Operators), Functions, and Greek Letters, as shown in Figure 3.



```
FUNCTION_VERBAL_DICT

{'sin': 'sine of',
 'cos': 'cosine of',
 'tan': 'tangent of',
 'log': 'logarithm of',
 'ln': 'natural logarithm of',
 'exp': 'exponential of'}
```

```
MATH_VERBAL_DICT

{'π': 'pi',
 '+': 'plus',
 '-': 'minus',
 '*': 'times',
 '/': 'divided by',
 '=': 'equals',
 '^': 'raised to the power of',
 '**': 'raised to the power of',
 '√': 'square root of'}
```

```
GREEK_VERBAL_DICT

{'α': 'alpha',
 'β': 'beta',
 'γ': 'gamma',
 'δ': 'delta',
 'θ': 'theta',
 'λ': 'lambda',
 'μ': 'mu',
 'σ': 'sigma',
 'Ω': 'omega',
 'ε': 'epsilon'}
```

Figure 3. Math dictionaries

Sentence-level context and neighbouring lexical clues are used to make the spoken math fit perfectly into the first sentence. Also addresses disambiguation (e.g., reading f(x) as "function f of x" rather than "f times x"). Finally, reinsert the verbalised phrases into the original document text, overwriting or adding the identified phrases while maintaining the original formatting and semantics. The transformed final document maintains logical coherence and is equivalent to the input in coverage. These sentences can be inline, block-level, or tag-encoded based on the document structure.

Mathematical Description of the CARMEN Model

Let $D = \{d_1, d_2, \ldots, d_n\}$be the set of documents containing mathematical expressions. Each document $d_i$consists of textual and symbolic parts:

$$d_i = T_i + M_i$$

where $T_i$is the text content and $M_i$represents the mathematical expressions.

The CARMEN model applies a **rule-based transformation function** $\mathcal{F}$ that normalises and verbalises each mathematical expression based on contextual semantics:

where $C_i$denotes the contextual information in the surrounding sentence.

The transformation process is sequential:

$$\mathcal{F} = f_3(f_2(f_1(M_i)))$$

with

- f_1: Regex-based mathematical expression detection
- f_2: Symbolic simplification using LaTeX2Text patterns
- f_3: Semantically-based context-verbalisation.

The reconstructed document is obtained as:

$$R_i = T_i + V_i$$

Quality of summing up the harmonic mean of ROUGE measures quantifies the quality of the summing up:

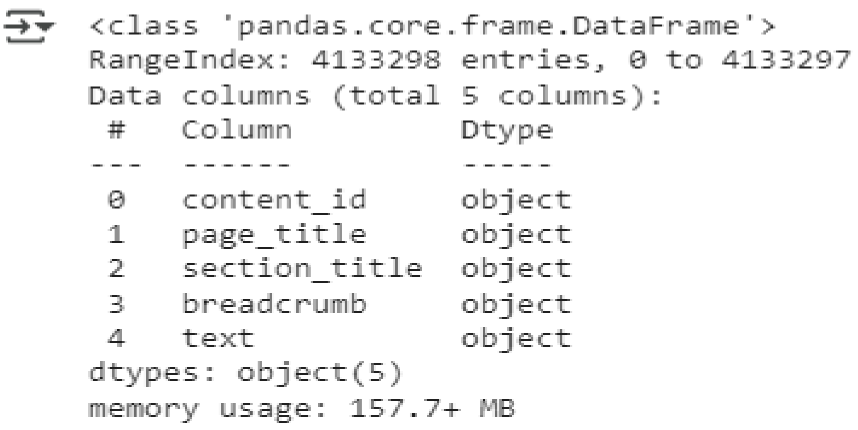
$$R = \frac{1}{3}(R_1 + R_2 + R_L)$$

where R_1, R_2, R_Ldenote ROUGE-1, ROUGE-2, and ROUGE-L scores respectively. The mathematical model is designed to preserve the symbolic content of mathematics and convert it into readable, context-sensitive text without semantic loss.

RESULTS AND DISCUSSION

Dataset

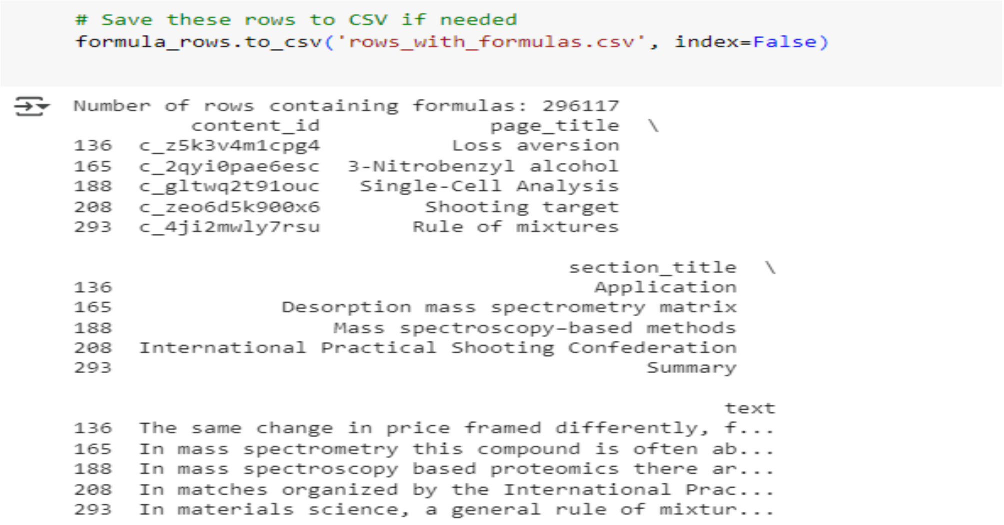
The proposed model is trained and analysed using the Kaggle dataset and the Wiki STEM Corpus (2024). This dataset is a STEM (Science, Technology, Engineering, and Mathematics) corpus, filtered by Wikipedia article category metadata [18]. In extracting the wiki page contents, we alleviated the common rendering problems (numbers, equations, and symbols) of current wiki datasets. Wikipedia mathematics articles are used to evaluate the proposed model's performance. These Wikipedia articles cover issues in mathematics and include formulae, equations, and technical notation. It supports inline and block-level mathematical expressions in either standard LaTeX or MathML. The reason is that human-readable explanations are provided in sections that help align math with natural language (training and assessment). To retrieve page content, the Wikipedia API Python module is used. It will be developed in Python using libraries such as sympy, PyMuPDF, LaTeX2Text, and the re module to extract LaTeX math regular expressions.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4133298 entries, 0 to 4133297
Data columns (total 5 columns):
 #   Column         Dtype
---  ------         -----
 0   content_id     object
 1   page_title     object
 2   section_title  object
 3   breadcrumb     object
 4   text           object
dtypes: object(5)
memory usage: 157.7+ MB
```

Figure 4. Raw dataset

The dataset contains 400,000 rows and five columns, as shown in the figure 4. From the given dataset, identify the rows with equations or formulas and create a separate dataset as shown in the figure 5. The new dataset consists of 2 lakh rows.



```
# Save these rows to CSV if needed
formula_rows.to_csv('rows_with_formulas.csv', index=False)
```

```
Number of rows containing formulas: 296117
         content_id                page_title  \
136  c_z5k3v4m1cpg4             Loss aversion
165  c_2qyi0pae6esc     3-Nitrobenzyl alcohol
188  c_gltwq2t91ouc     Single-Cell Analysis
208  c_zeo6d5k900x6           Shooting target
293  c_4ji2mwly7rsu          Rule of mixtures

                                  section_title  \
136                                 Application
165            Desorption mass spectrometry matrix
188                Mass spectroscopy-based methods
208  International Practical Shooting Confederation
293                                     Summary

                                          text
136  The same change in price framed differently, f...
165  In mass spectrometry this compound is often ab...
188  In mass spectroscopy based proteomics there ar...
208  In matches organized by the International Prac...
293  In materials science, a general rule of mixtur...
```

Figure 5. Final dataset

Performance Measures

The proposed model's performance is evaluated using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores, including ROUGE-1, ROUGE-2, and ROUGE-L. ROUGE 1 computes the overlap for unigrams (one word), which is computed using Equation 4.1. ROUGE-2 computes the overlap for bigrams (two-word sequences) using Equation 4.2 (Raghav et al., 2022). The n-gram approachiseasyandinexpensivetocompute,sotheapproachhasbeenappliedextensively to summary evaluation [17].ROUGE-L computes the length of the longest common subsequence of the reference and system summaries [14], which is calculated using the Equ. 4.3.

$$\left.\begin{aligned} ROUGE1-Recall &= \frac{w\in Reference^{minCount_{cand}(W),Count_{ref}(W)}}{w\in Reference^{Count_{ref}(W)}} \\ ROUGE1-Precision &= \frac{w\in Reference^{minCount_{cand}(W),Count_{ref}(W)}}{w\in Candidate^{Count_{cand}(W)}} \\ ROUGE1-F1\ Score &= \frac{2.Precision.Recall}{Precision+Recall} \end{aligned}\right\}\text{--- Equ. (4.1)}$$

$$\left.\begin{aligned} ROUGE2-Recall &= \frac{bg\in Reference^{minCount_{cand}(bg),Count_{ref}(bg)}}{bg\in Reference^{Count_{ref}(bg)}} \\ ROUGE2-Precision &= \frac{bg\in Reference^{minCount_{cand}(bg),Count_{ref}(bg)}}{bg\in Candidate^{Count_{cand}(bg)}} \\ ROUGE2-F1\;Score &= \frac{2.Precision.Recall}{Precision+Recall} \end{aligned}\right\}\text{--- Equ. (4.2)}$$

Where w referred to a word, bg referred to a bigram,$Count_{cand}$ and $Count_{ref}$Refers to the number of times a word appeared in the candidate and the reference, respectively.

$$\left.\begin{aligned} R_{LCS} &= \frac{LCS(C,R)}{n} \\ P_{LCS} &= \frac{LCS(C,R)}{m} \\ ROUGEL-F1\;Score &= \frac{(1+\beta^2).P_{LCS}\;\;.\;\;R_{LCS}}{R_{LCS}+\beta^2.P_{LCS}} \end{aligned}\right\}\text{---}\qquad\qquad\text{Equ. (4.3)}$$

Where LCS referred to Length of Common Subsequence between Candidate (C) and Reference (R), m is Length of Candidate, n is Length of Reference, β is 1 for a balanced F1–score, and R_{LCS} *and* P_{LCS}Are Recall and Precision, respectively. ROUGE-L estimates the length of the longest common subsequence between the system summary and the reference summaries [15]. ROUGE-L is particularly important for assessing the coherence and fluency of abstracts, as it identifies word-order patterns. To perform validation of ROUGE-1, ROUGE-2, and ROUGE-L, the mathematical expression and its verbalised output (as generated by the model) are compared to a human-written reference. These ROUGE scores calculate the Harmonic mean of precision and recall. The proposed model, compared with previous studies, highlighted MathSpeak [31] and W3C MathML [33] Verbalizers.

Table 3. Evaluation of models based on ROUGE score

Methods	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Mathspeak	0.661290	0.568966	0.661290
W3C MathML	0.710227	0.357143	0.710227
CARMEN (Proposed)	0.863636	0.833333	0.863636

Table 3 presents the performance evaluation of the proposed model, CARMEN,based on ROUGE scores. All proposed CARMEN has ROUGE > 0.8, which is exceptional in natural language tasks [29]. Moreover, it scores highest on ROUGE-1 and ROUGE-L, implying it is nearly identical to the reference in terms of words used and structure. CARMEN achieves higher ROUGE-2 scores than the other two models, suggesting more fluent phrase construction, despite a slightly lower vocabulary overlap. These scores indicate that model CARMEN is performing at a very high level in both content fidelity and summary fluency.

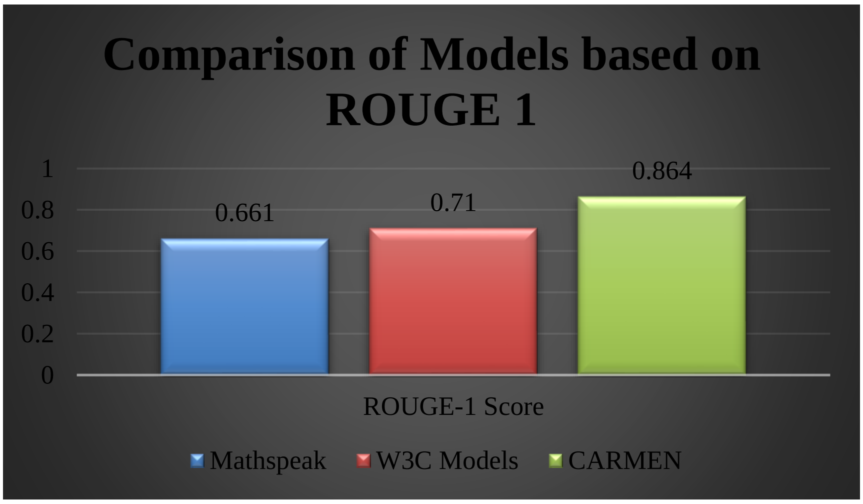


Figure 6. Comparison of models based on ROUGE 1

Figure 6 illustrates a pictorial representation of the proposed CARMEN model based on the ROUGE-1 metric. The proposed model achieves the highest ROUGE-1 score of 0.863636, whereas Mathspeak and W3c MathML verbalizer score 0.661290 and 0.710227, respectively. From this, it is clear that the CARMEN model is more reliable in referencing terms and in structuring sentences.

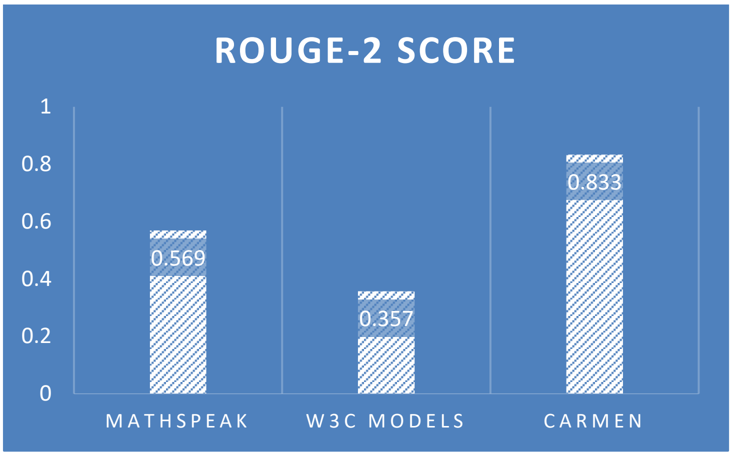


Figure 7. Comparison of models based on ROUGE 2

Figure 7 illustrates a pictorial representation of the proposed CARMEN model based on the ROUGE-2 metric. The proposed model achieves the highest ROUGE-1 score of 0.83333, whereas Mathspeak and W3c MathML verbalizer score 0.56966 and 0.357143, respectively. From this, it is clear that the CARMEN model is more fluent in phase construction.

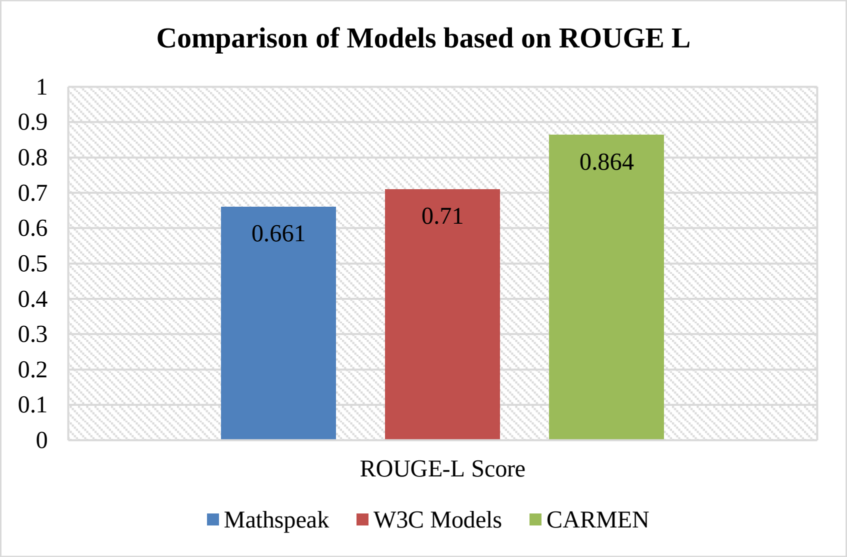


Figure 8. Comparison of models based on ROUGE L

Figure 8 illustrates a pictorial representation of the proposed CARMEN model based on the ROUGE-L metric. The proposed model achieves the highest ROUGE-L score of 0.863636, whereas Mathspeak and the W3C MathML verbalizer score 0.661290 and 0.710227, respectively. From this, it is clear that the CARMEN model is more reliable in referencing terms and in framing sentence structure.

CONCLUSION

The majority of existing models of summarisation assume that mathematical content can be ignored or manipulated, and that it can be dropped or distorted. CARMEN is a Rule-Based Math Expression Normalisation and Verbalisation that represents symbolic equations in readable natural language using domain-specific rules [26]. It is not easily tempted into extractive models (redundancy, inflexibility) or abstractive models (hallucination, information loss) to give out summaries that are linguistically sound but factual [23][24][25]. The ROUGE-based CARMEN evaluation demonstrates its efficiency in generating high-quality, human-like summaries. The model alleviates the balance between lexical fidelity and narrative coherence, which is why it can be applied most effectively to real-world deployment in educational, legal, and medical document summarisation tasks. The results make CARMEN a strong, sound model for hybrid summarisation. A new contribution of the CARMEN model is an architecture that enables synergistic, rule-based simplification, preserves semantic context, and supports mathematical verbalisation. It provides significant methodological improvements over current models. Its rule-laden, context-sensitive system not only enhances the fidelity and fluency of content but also uniquely handles technical material, such as mathematical notation. The mentioned benefits make it an excellent option as a central system for summarising academic, scientific, and legal documents. Most summary systems have been trained and adapted to general text (i.e. newspaper articles), CARMEN is specifically designed to process multimodal text, such as scientific PDFs, technical manuals, and academic text containing embedded symbols and equations. The proposed model performs well, producing high-quality summaries, as indicated by the ROUGE scores from the trials. The findings confirm the model's effectiveness in producing well-organised, coherent, and content-based summaries. The model's reliance on consistent performance across trials gives it reliability as a summarisation tool, capable of producing high-quality output for real-time applications [30]. The model consistently performs well, achieving high ROUGE-1 and ROUGE-L scores of 0.863636, correctly selecting the essential information and the primary content of the source documents.

REFERENCES

[1] Mukhiddinov M, Kim SY. A systematic literature review on the automatic creation of tactile graphics for the blind and visually impaired. Processes. 2021 Sep 26;9(10):1726. https://doi.org/10.3390/pr9101726

[2] Aguinis H, Gottfredson RK, Joo H. Best-practice recommendations for defining, identifying, and handling outliers. Organizational research methods. 2013 Apr;16(2):270-301. https://doi.org/10.1177/1094428112470848

[3] Rahimi I, Gandomi AH, Chen F, Mezura-Montes E. A review on constraint handling techniques for population-based algorithms: from single-objective to multi-objective optimization. Archives of Computational Methods in Engineering. 2023 Apr;30(3):2181-209. https://doi.org/10.1007/s11831-022-09859-9

[4] Wang S, Cheah JH, Wong CY, Ramayah T. Progress in partial least squares structural equation modeling use in logistics and supply chain management in the last decade: a structured literature review. International Journal of Physical Distribution & Logistics Management. 2024 Oct 17;54(7/8):673-704. https://doi.org/10.1108/ijpdlm-06-2023-0200

[5] Psomas E. Future research methodologies of lean manufacturing: a systematic literature review. International Journal of Lean Six Sigma. 2021 Nov 19;12(6):1146-83. https://doi.org/10.1108/ijlss-06-2020-0082

[6] Zhu X, Zhang G, Sun B. A comprehensive literature review of the demand forecasting methods of emergency resources from the perspective of artificial intelligence. Natural Hazards. 2019 May 15;97(1):65-82. https://doi.org/10.1007/s11069-019-03626-z

[7] Fiandrino S, Tonelli A, Devalle A. Sustainability materiality research: a systematic literature review of methods, theories and academic themes. Qualitative Research in Accounting & Management. 2022 Oct 21;19(5):665-95. https://doi.org/10.1108/qram-07-2021-0141

[8] Liguda C, Pfeiffer T. Modeling math word problems with augmented semantic networks. In International Conference on Application of Natural Language to Information Systems 2012 Jun 26 (pp. 247-252). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31178-9_29

[9] Zhou Q, Huang D. Towards generating math word problems from equations and topics. In Proceedings of the 12th international conference on natural language generation 2019 (pp. 494-503). https://doi.org/10.18653/v1/w19-8661

[10] Chai CP. Comparison of text preprocessing methods. Natural language engineering. 2023 May;29(3):509-53. https://doi.org/10.1017/s1351324922000213

[11] Amato A, Di Lecce V. Data preprocessing impact on machine learning algorithm performance. Open computer science. 2023 Jul 17;13(1):20220278. https://doi.org/10.1515/comp-2022-0278

[12] Helin R, Indahl UG, Tomic O, Liland KH. On the possible benefits of deep learning for spectral preprocessing. Journal of Chemometrics. 2022 Feb;36(2): e3374. https://doi.org/10.1002/cem.3374

[13] Feichter C, Schlippe T. Investigating models for the transcription of mathematical formulas in images. Applied Sciences. 2024 Jan 29;14(3):1140. https://doi.org/10.3390/app14031140

[14] Peyrard M, Eckle-Kohler J. Optimizing an approximation of rouge-a problem-reduction approach to extractive multi-document summarization. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2016 Aug (pp. 1825-1836). https://doi.org/10.18653/V1/P16-1172

[15] Lee D, Shin MC, Whang T, Cho S, Ko B, Lee D, Kim E, Jo J. Reference and document aware semantic evaluation methods for Korean language summarization. InProceedings of the 28th International Conference on Computational Linguistics 2020 Dec (pp. 5604-5616). https://doi.org/10.18653/v1/2020.coling-main.491

[16] Jain R, Mavi V, Jangra A, Saha S. Widar-weighted input document augmented rouge. InEuropean Conference on Information Retrieval 2022 Apr 5 (pp. 304-321). Cham: Springer International Publishing. https://doi.org/10.48550/arxiv.2201.09282

[17] Ganesan K. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. arXiv preprint arXiv:1803.01937. 2018 Mar 5. https://arxiv.org/abs/1803.01937

[18] Biswas R, Bamba U, Broad N. Wiki STEM Corpus. Kaggle; 2024 Apr 12.

[19] Latif A, Kim J. Evaluation and analysis of large language models for clinical text augmentation and generation. IEEE Access. 2024 Apr 3; 12:48987-96. https://doi.org/10.1109/ACCESS.2024.3384496

[20] Muludi K, Fitria KM, Triloka J. Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model. International Journal of Advanced Computer Science & Applications. 2024 Mar 1;15(3). https://doi.org/10.14569/ijacsa.2024.0150379

[21] Chai Y, Xie H, Qin JS. Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities. Artificial Intelligence Review. 2026 Jan;59(1):35.

[22] Wang JH, Zhuang JA. Text Summary Generation based on Data Augmentation and Contrastive Learning. In2024 IEEE International Conference on Big Data (BigData) 2024 Dec 15 (pp. 7218-7224). IEEE. https://doi.org/10.1109/BigData62323.2024.10825107

[23] Alsultan R, Sagheer A, Hamdoun H, Alshamlan L, Alfadhli L. PEGASUS-XL with saliency-guided scoring and long-input encoding for multi-document abstractive summarization. Scientific Reports. 2025 Jul 22;15(1):26529.

[24] Jia R, Zhang X, Cao Y, Lin Z, Wang S, Wei F. Neural label search for zero-shot multi-lingual extractive summarization. InProceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2022 May (pp. 561-570). https://doi.org/10.18653/v1/2022.acl-long.42

[25] Syed AA, Gaol FL, Boediman A, Matsuo T, Budiharto W. A survey of abstractive text summarization utilising pretrained language models. InAsian Conference on Intelligent Information and Database Systems 2022 Nov 28 (pp. 532-544). Cham: Springer International Publishing.

[26] Guo B, Gong Y, Shen Y, Han S, Huang H, Duan N, Chen W. Genius: Sketch-based language model pre-training via extreme and selective masking for text generation and augmentation. arXiv preprint arXiv:2211.10330. 2022 Nov 18.

[27] Sun D, Lyu Y, Li J, Liu X, Kara D, Lebiere C, Abdelzaher T. The Irrational LLM: Implementing Cognitive Agents with Weighted Retrieval-Augmented Generation. In 2025 34th International Conference on Computer Communications and Networks (ICCCN) 2025 Aug 4 (pp. 1-9). IEEE. https://doi.org/10.1109/ICCCN65249.2025.11134012

[28] Phang J, Zhao Y, Liu PJ. Investigating efficiently extending transformers for long input summarization. InProceedings of the 2023 Conference on Empirical Methods in Natural Language Processing 2023 Dec (pp. 3946-3961). https://doi.org/10.18653/v1/2023.emnlp-main.240

[29] Corallo G, Papotti P. Finch: Prompt-guided key-value cache compression for large language models. Transactions of the Association for Computational Linguistics. 2024 Nov 18; 12:1517-32. https://doi.org/10.1162/tacl_a_00716

[30] Guo Z, Adedigba AP, Mallipeddi R. Cluster-Aggregated Transformer: Enhancing lightweight parameter models. Engineering Applications of Artificial Intelligence. 2025 Nov 1; 159:111468. https://doi.org/10.1016/j.engappai.2025.111468

[31] Asebriy Z, Raghay S, Bencharef O. An assistive technology for braille users to support mathematical learning: a semantic retrieval system. Symmetry. 2018 Oct 26;10(11):547. https://doi.org/10.3390/sym10110547

[32] Rotard M, Knödler S, Ertl T. A tactile web browser for the visually disabled. InProceedings of the sixteenth ACM conference on Hypertext and hypermedia 2005 Sep 6 (pp. 15-22). https://doi.org/10.1145/1083356.1083361

[33] Tian X, Wang J. Retrieval of scientific documents based on HFS and BERT. IEEE Access. 2021 Jan 5; 9:8708-17. https://doi.org/10.1109/ACCESS.2021.3049391