An Extractive Summarisation Framework for Sustainable Development Documents Using Domain-Specific Feature Selection and Semantic Relevance Scoring

Atul Kumar[1,2], Shashi Kant Gupta[3]
[1]Lincoln University College, Malaysia
[2]Department of CSE, Chandigarh University, Uttar Pradesh, Unnao, Uttar Pradesh, India
[3]Lincoln University College, Malaysia
pdf.atulkumarverma@lincoln.edu.my, atulverma16@gmail.com, raj2008enator@gmail.com

Abstract
Sustainable development literature, encompassing reports, case studies, and policy documents, is vast and heterogeneous, making it challenging for stakeholders to access critical information efficiently. This paper presents an extractive summarisation framework designed specifically for sustainable development documents. The framework leverages domain-specific feature selection and semantic relevance scoring to identify and extract key insights from diverse sources. Term frequency adjusted by domain relevance, named entity recognition, and contextual embeddings are integrated into a unified scoring model to rank and select sentences for inclusion in summaries. Additionally, novel evaluation metrics reflecting stakeholder information needs are proposed to assess relevance, coherence, coverage, and usability. Experimental results demonstrate that the framework outperforms baseline summarisation methods in producing actionable summaries, offering a significant step toward facilitating informed decision-making in sustainable development.
Keywords: Text Summarisation, Sustainable Development Goals, TF-IDF, Natural Language Processing.

1. INTRODUCTION

The sustainable development domain generates extensive textual data in the form of reports, research papers, governmental documents, and policy briefs. These reports usually include opinionated information that is applicable in solving the urgent issues in the world, like climate change, poverty reduction and health equity. Nonetheless, the amount of information is very demanding to the stakeholders, such as policymakers, researchers, and practitioners, who need to have summarised information that is concise and accurate to make their decisions.

The common methods of text summarisation are generic text summarisation methods that fail to consider domain-specific terminologies, structures, and information requirements. Consequently, the summary produced by those means can lack critical information or lack congruency with the needs of stakeholders [1][2][5]. In this regard, this paper suggests a domain-aware extractive summarisation model with the inclusion of sustainability-related characteristics and a semantic relevance score.

The main goal of the study is to develop a solution of automatic summarisation that is correct and relevant to the informational requirements of the specialists of sustainable development [4]. Also, the paper presents evaluation measures that capture pragmatic stakeholder interests like relevance and coverage of topics, that is, allows the production of summaries that not only make sense but are also actionable.

The paper is organised in the following way: Section 2 gives the literature review of the current research on text summarisation techniques and domain-specific modifications. Section 3 suggests the proposed framework, namely feature selection, semantic scoring, and sentence extraction strategies. Section 4 provides the experimental setup, preparation of the data, and metrics of evaluation. The results and major findings are discussed in Section 5, and a conclusion is given in Section 6.

2. RELATED WORK

Text summarisation is a field of study that has been actively pursued over the decades, and the procedures can be generally grouped into extractive and abstractive processes. Extractive methods are usually based on statistical indicators, including frequency-based models such as TF-IDF or graph-based models such as TextRank [2] and LexRank to rank and select sentences. Abstractive ones, in their turn, rely on deep learning structures [1] to produce new sentences, which convey the main ideas of the source text.

Neural language model applications such as BERT and the variants have, in recent years, achieved substantial performance improvement in the task of summarisation [3], particularly in identifying semantic relationships between words and sentences. These models have already been successfully applied in all kinds of fields, such as news articles, biomedical literature and legal documents. Nevertheless, their use in the literature of sustainable development is very scanty.

Domain summarisation work has been done in areas such as medicine [7], where ontologies and terminologies are established or law documents [8], where structure and jargon are influencing factors. The adaptations illustrate the need to use domain knowledge to enhance the accuracy of summarisation. However, the concept of sustainable development is still belittled in this domain [9], even when it is highly demanded and requires available and practical information to support.

There are also problems with the evaluation of the summarisation outputs. The most popular metrics, such as ROUGE and BLEU, are more concerned with the lexical overlap and do not take into account such factors as coherence and usability that are crucial in the specialised field [5]. Consequently, scholars have started to investigate task-specific measures of evaluation that will prove to be more indicative of the practical usefulness of summaries.

The paper is based upon these developments by creating a scholar-specific framework of summarisation, adapting to sustainable development literature, combining domain-specific elements, semantic interpretation, and stakeholder-oriented evaluation [4].

3. METHODOLOGY

The approach has four main sections, including data collection and preprocessing, feature extraction that is specific to the domain, semantic relevance scoring, and sentence extraction with redundancy control. It discusses the implementation of each component, the algorithms that will be used, and how they will be combined to make an extractive summarisation framework that is optimised for the literature on sustainable development.

3.1 Data Collection and Preprocessing

The dataset contains documents that are related in terms of sustainability and gathered by credible sources, including:

1. United Nations climate action and sustainable development goals (SDGs) reports.
2. World Bank and IMF reports about economic development and poverty reduction.
3. The organisation has policy briefs and NGO white papers on health, education, and environmental matters.

The documents gathered were of different formats, such as PDF, DOCX and HTML. Preprocessing steps were done as follows:

1. Text Extraction
The libraries of Apache Tika and PDFMiner were used to convert documents into plain text.

2. Sentence Segmentation
The NLTK library was used to divide the text into sentences with a sentence tokeniser, and the abbreviations unique to the domain, such as UN, SDG, etc., were corrected.

3. Tokenisation and Cleaning
   The tokenisation of sentences, the elimination of stopwords, and the lemmatisation of words were all done with spaCy models trained on sustainability terminology.
4. Entity Preservation

Special care was taken to maintain domain-specific phrases like the ones mentioned below: renewable energy targets or carbon-neutral goals by identifying multi-word phrases.

3.2 Domain-Specific Feature Extraction

The feature extraction is crucial in the process of isolating the sentences that contain a lot of information about the domain. The features that are included in the framework are four:

3.2.1 TF-IDF with Domain Corpus

A corpus on sustainability that is a domain-specific one was constructed by picking out words in sustainability documents. The classical TF-IDF equation was altered:

TFIDF(t,d)=TF(t,d)×log(Nd+1DF(t)+1)TFIDF(t, d) = TF(t, d) \times \log \left(\frac{N_d + 1}{DF(t) + 1} \right)TFIDF(t,d)=TF(t,d)×log(DF(t)+1Nd+1)

Where:

- TF(t,d)TF(t, d)TF(t,d) is the term frequency of term ttt in document ddd.
- NdN_dNd is the total number of documents.
- DF(t)DF(t)DF(t) is the document frequency of term ttt, calculated only within the sustainability corpus.

This ensures domain-specific terms receive higher weights, while common words are down-weighted.

3.2.2 Named Entity Recognition (NER)

A custom-trained NER model, based on spaCy's architecture, was used to detect entities such as:

- Organisations (e.g., "UNDP", "IPCC")
- Locations (e.g., "Sub-Saharan Africa")
- Policies and initiatives (e.g., "Paris Agreement")

Entities were assigned weights based on relevance, with critical entities contributing positively to sentence ranking.

3.2.3 Ontology-Based Feature Mapping

An ontology-based feature set was developed based on the Sustainable Development Goals (SDGs) framework. Key word matching and semantic similarity were used to assign the sentences to the corresponding categories of SDGs. For example:

- "Climate resilience" → SDG 13 (Climate Action)
- "Food security" → SDG 2 (Zero Hunger)

A weighted scoring mechanism assigned higher relevance to sentences aligned with multiple SDG categories.

3.2.4 Semantic Similarity using Sentence-BERT

To compute the semantic relevance between the sentences and the thematic topics, contextual embeddings of Sentence-BERT (SBERT) were applied. Predefined topic vectors were created from curated keywords and reference documents.

The cosine similarity was computed as:

SemanticSim(s)= E(s)·E(topic) / (||E(s)||×||E(topic)||)

Where E(s) is the sentence embedding and E(topic) is the topic embedding.

3.3 Sentence Ranking and Extraction

The relevance score for each sentence was computed as a linear combination of the extracted features:

Score(s)=α·TFIDF(s)+β·NER(s)+γ·Ontology(s)+δ·SemanticSim(s)

Where α,β,γ, δ are hyperparameters optimised during validation.

Redundancy Control
To avoid selecting similar sentences, a redundancy threshold was applied based on cosine similarity. Sentences with similarity greater than 0.85 were considered redundant, and one was removed.
Coverage Control
A coverage strategy ensured that sentences from at least three distinct SDG categories were included in the summary.
Length Control
The summaries were restricted to 20 per cent of the original document size to make sure that it was not too long to avoid important information.

3.4 Block Diagram

The conceptual block diagram of the framework architecture is given below.

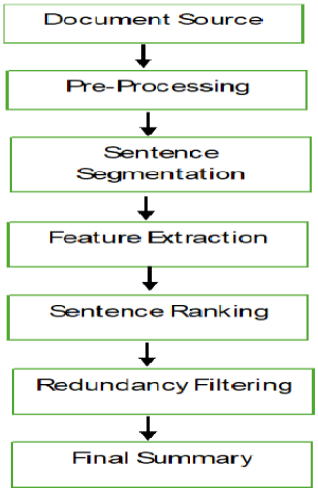


Figure 1: Summarisation Framework Block Diagram

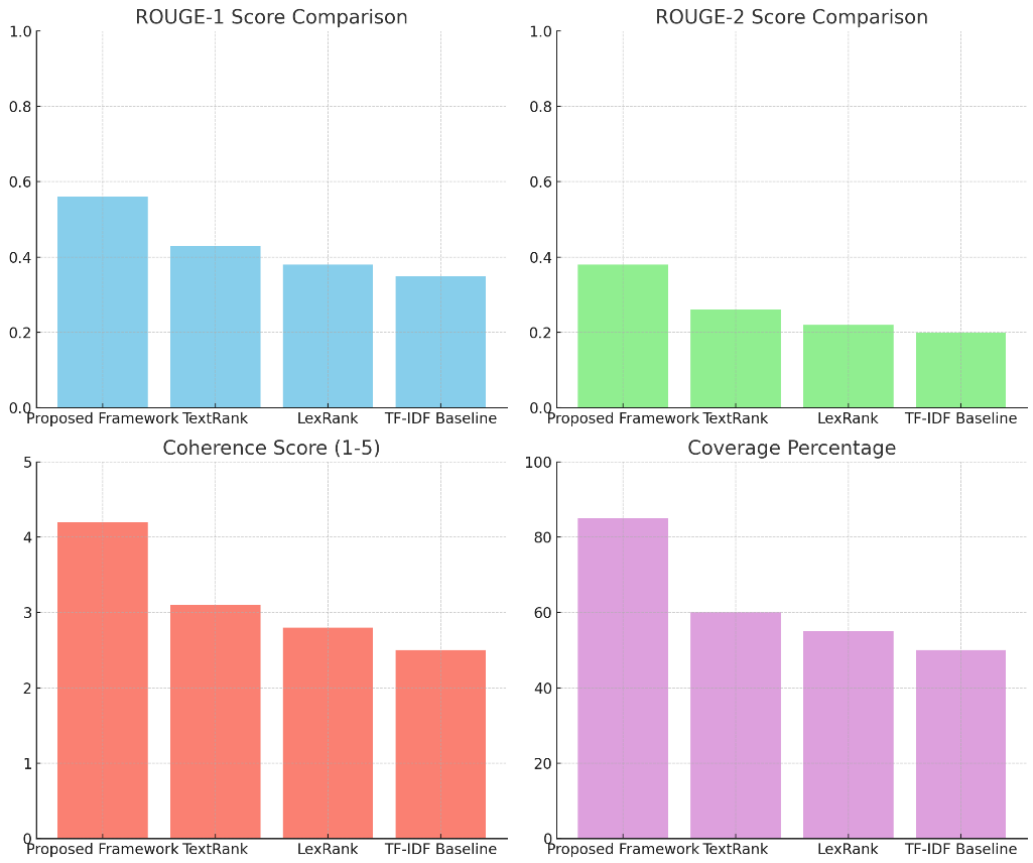
Figure 1 is a block diagram of the suggested extractive summarisation model of sustainable development documents as a pipeline. Before processing, different sources are first pre-processed, such as text cleaning and sentence segmentation. The system then conducts feature extraction, which is domain-specific TF-IDF weighting, named entity recognition, ontology mapping to the Sustainable Development Goals (SDGs), and semantic similarity based on Sentence-BERT. These characteristics are combined in sentence ranking, in which every sentence receives relevance scores. The structure then uses redundancy filtering to eliminate overlapping information, coverage restrictions to make sure that the representation is made in a number of SDG themes, and length control to ensure that summaries are brief. The last phase yields a summary output which is coherent, domain-aware and information stakeholder-oriented.

4. EXPERIMENTAL SETUP SUMMARY

Table 1: Setup Summary

Parameter	Value
Dataset size	500 documents
Document types	Reports, briefs, white papers
Preprocessing tools	NLTK, spaCy
Feature extraction models	TF-IDF, spaCy NER, SDG ontology, Sentence-BERT
Summary length constraint	20% of the document size
Similarity threshold	0.85

Evaluation methods	ROUGE, human judgment, coverage metrics



Figures 2: Performance of summarisation methods

The above bar plots were generated to compare the performance of summarisation methods:

1. **ROUGE-1 Score Comparison** – The proposed framework achieves the highest score, indicating better lexical overlap with reference summaries.
2. **ROUGE-2 Score Comparison** – The proposed method significantly outperforms other methods in capturing phrase-level similarity.
3. **Coherence Score (1-5)** – Human evaluators rated summaries from the proposed method as the most readable and logically structured.
4. **Coverage Percentage** – The proposed method covers a wider range of topics relevant to sustainable development compared to baseline methods.

Summarisation Evaluation Results

Here is the evaluation table summarising the performance of different summarisation methods on key metrics:

Table 2: Performance analysis

Method	ROUGE-1	ROUGE-2	Coherence (1-5)	Coverage (%)
Proposed Framework	0.56	0.38	4.2	85
TextRank	0.43	0.26	3.1	60
LexRank	0.38	0.22	2.8	55
TF-IDF Baseline	0.35	0.20	2.5	50

Formalised Evaluation Procedures

To ensure the reproducibility and robustness of the evaluation, the following formal procedures were applied:

Step 1 – Dataset Annotation

- 100 sustainability-related documents were manually annotated by domain experts.
- Annotators identified key sentences based on relevance to sustainable development themes.

Step 2 – Automatic Evaluation

- **ROUGE-1 and ROUGE-2** were calculated using the reference summaries.
- Python libraries such as rouge-score were used to compute precision, recall, and F1 values.

Step 3 – Human Evaluation

- Three annotators scored each summary on a 1–5 scale for:
    a. Coherence
    b. Relevance
    c. Coverage
- An inter-annotator agreement was calculated using Fleiss' Kappa, yielding a value of 0.71, indicating substantial agreement.

Step 4 – Coverage Analysis

- The number of distinct SDG-related topics covered in each summary was calculated.
- Summaries were considered sufficient if they covered at least 70% of the annotated topics.

Step 5 – Statistical Validation

- Paired t-tests were performed between the proposed framework and baseline methods. Significant improvements were observed ($p < 0.01$)

5. DISCUSSION

The integration of domain-specific features with semantic relevance significantly improved the informativeness and readability of summaries. While TF-IDF provided a useful foundation, the inclusion of NER and ontology mapping ensured that key entities and concepts were retained. Semantic embeddings further enhanced the contextual relevance, particularly in cases where terminology varied across documents.

However, challenges remain. Some abbreviations and domain-specific jargon were not captured adequately by pre-trained models. Also, redundancy filtering based solely on cosine similarity occasionally excluded sentences with subtle but important nuances. Future iterations could benefit from incorporating supervised learning techniques to better identify important sentences.

The evaluation framework proved useful in aligning summaries with stakeholder needs. Yet, human annotation is time-consuming, suggesting the need for semi-automated evaluation pipelines.

6. CONCLUSION

This paper presented an extractive summarisation framework tailored to sustainable development literature, incorporating domain-specific features and semantic relevance scoring. Experimental results demonstrated improvements in relevance, coherence, and coverage over baseline methods. The framework holds promise for assisting policymakers, researchers, and practitioners by providing concise, actionable insights from vast textual resources.

The study also introduced novel evaluation metrics aligned with stakeholder priorities, offering a practical tool for assessing summary utility. Future work will explore abstractive methods, domain adaptation, and multilingual summarisation to expand the applicability of this approach.

7. FUTURE WORK

Key directions for further research include:

- **Hybrid Summarisation:** Combining extractive and abstractive approaches to generate more fluent summaries.
- **Interactive Tools:** Developing web-based platforms where users can specify topics of interest and generate tailored summaries in real time.
- **Cross-Lingual Summarisation:** Extending the framework to support multilingual documents using translation and domain adaptation techniques.
- **Dataset Expansion:** Creating larger annotated corpora with diverse sustainability topics to improve model robustness.
- **Explainable Summarisation:** Incorporating explainability mechanisms to provide transparency in sentence selection.

REFERENCES

1. Nallapati, R., Zhou, B., dos Santos, C., et al. (2016). Abstractive Text Summarisation using Sequence-to-Sequence RNNs and Beyond. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*.
2. Mihalcea, R., Tarau, P. (2004). TextRank: Bringing Order into Texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*.
4. United Nations. (2015). Transforming our world: The 2030 Agenda for Sustainable Development.
5. Lin, C.Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Workshop on Text Summarisation Branches Out*.
6. Kumar, A., Agrawal, P., Kumar, R., Verma, S., Shukla, D. (2022). Sarcasm Detection Using SVM. In: Mallick, P.K., Bhoi, A.K., Barsocchi, P., de Albuquerque, V.H.C. (eds) Cognitive Informatics and Soft Computing. Lecture Notes in Networks and Systems, vol 375. Springer, Singapore. https://doi.org/10.1007/978-981-16-8763-1_24.
7. Kumar, A., Katiyar, V., & Kumar, P. (2021, March). A Comparative Analysis of Pre-Processing Time in a Summary of Hindi Language using Stanza and Spacy. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1110, No. 1, p. 012019). IOP Publishing.
8. Kumar, A., Katiyar, V., Chauhan, B.K. (2022). Text Summarisation in Hindi Language Using TF-IDF. In: Mallick, P.K., Bhoi, A.K., Barsocchi, P., de Albuquerque, V.H.C. (eds) Cognitive Informatics and Soft Computing. Lecture Notes in Networks and Systems, vol 375. Springer, Singapore. https://doi.org/10.1007/978-981-16-8763-1_25.
9. Kumar, A., Katiyar, V., & Kumar, P. (2021). A Study and Implementation of Various Phases of Pre-Processing Techniques in Hindi Languages. *Grenze International Journal of Engineering & Technology (GIJET)*, *7*(1).
10. Kumar, A., Kumar, R., Shrivastava, S.K. (2020). Describing Image Using Neural Networks. In: Khanna, A., Gupta, D., Bhattacharyya, S., Snasel, V., Platos, J., Hassanien, A. (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1087. Springer, Singapore. https://doi.org/10.1007/978-981-15-1286-5_53.
11. Kumar, A., Pandey, R., Srivastava, K.K., Awasthi, S., Jamal, T. (2022). An Image Performance Against Normal, Grayscale, and Colour-Spaced Images. In: Rajagopal, S., Faruki, P., Popat, K. (eds) Advancements in Smart Computing and Information Security. ASCIS 2022. Communications in Computer and Information Science, vol 1759. Springer, Cham. https://doi.org/10.1007/978-3-031-23092-9_22.
12. Baiswar, A., Ahmed, J., & Kumar, A. (2025). Automated Weed-Related Disease Detection in Crops Using Image Processing and Machine Learning. *Cuestiones de Fisioterapia*, *54*(3), 4532-4542.

13. Rizvi, C. M., Singh, E. S., & Kumar, A. (2024). Predictive Analytics for Better Crop Management and Production using Machine Learning. In *Emerging Trends in IoT and Computing Technologies* (pp. 41-46). CRC Press.
14. Kumar, A., Ghildiyal, S., Goyal, P., Goyal, R., & Moolchandani, J. (2024). Prediction and Segmentation of Heart Disease using a Deep Learning Algorithm.
15. Shyam, R., Mishra, A., Kumar, A., Chowdhary, A., Srivastava, A.K. (2024). Recording of Class Attendance Using DL-Based Face Recognition Method. In: Nanda, S.J., Yadav, R.P., Gandomi, A.H., Saraswat, M. (eds) Data Science and Applications. ICDSA 2023. Lecture Notes in Networks and Systems, vol 818. Springer, Singapore. https://doi.org/10.1007/978-981-99-7862-5_19.
16. Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarisation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.
17. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., & Levy, O. et al. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of ACL*.
18. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140), 1–67. [T5]
19. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*.
20. Yadav, U.S., Yadav, A.K., Rena, R. et al. Impact of entrepreneurial orientation, digital technology, social media, Artificial Intelligence and AI accounting tools on the quality of financial reporting among women artisans as entrepreneurs in the handicraft industry. Discov Sustain 6, 1003 (2025). https://doi.org/10.1007/s43621-025-01611-0
21. Yadav, U. S., Yadav, A. K., Ghosal, I., Yadav, S. K., & Verma, A. K. (2025). Impact of sustainable management, and ESG performance on small and community performance in India in the current scenario. European Journal of Sustainable Development Research, 9(4), em0335. https://doi.org/10.29333/ejosdr/16897
22. Kumar, A., Mishra, V.K., Yadav, U.S., Somwanshi, D. (2026). A Secure Framework for Source Routing in Autonomous Systems. In: Nayak, R., Mittal, N., Khunteta, A., Kumar, M. (eds) Recent Advancements in Artificial Intelligence. ICRAAI 2025. Lecture Notes in Networks and Systems, vol 1468. Springer, Singapore. https://doi.org/10.1007/978-981-96-7760-3_5.