From Inconsistent Descriptions to Verified Retrieval Through a Local Data Cleaning Framework for Higher Education Climate Action Keyword Mapping

PMPC Gunathilake[1*], Tilak Hewawasam [2], Jagath Gunatilake [3]

[1] Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka

[2] Department of Geography, University of Peradeniya, Peradeniya, Sri Lanka

[3] Department of Geology, University of Peradeniya, Peradeniya, Sri Lanka

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.1026EDU0324

Received: 18 May 2026; Accepted: 23 May 2026; Published: 13 June 2026

ABSTRACT

Higher Education Institutions (HEIs) are increasingly expected to report sustainability and climate-related activities aligned with Nationally Determined Contributions (NDCs), aggregated through Locally Determined Contributions (LDCs), and recognised within national climate governance systems. This alignment is difficult when university activity descriptions use incomplete, inconsistent, or non-standard climate keywords, preventing automated detection of climate actions and leading to incorrect NDC mapping and unreliable inputs for future carbon quantification. NDC documents further compound this challenge by containing structurally similar actions across different sectors and contradictory actions arising from sectoral trade-offs, issues that often stem from limited cross-sectoral coordination during preparation. This study develops a data cleaning framework for higher education climate action keyword mapping, using Sri Lanka as a pilot study. Sri Lanka's updated NDCs serve as the national reference framework, enabling the methodology to be tested within a context where climate reporting infrastructure and cross-sectoral coordination remain emergent. The methodology applies Natural Language Processing (NLP) techniques, Jaccard similarity, and cosine similarity to identify sectoral overlaps and action-level conflicts. The framework then cleans climate-policy and sustainability-reporting text by removing duplicates, normalising terminology, preserving metadata, and constructing a locally relevant keyword set for mitigation, adaptation, and cross-cutting climate actions derived from sources aligned with IPCC, UNFCCC, NDC, SDG, and university sustainability reporting standards. Pilot results from Sri Lanka confirm the framework's effectiveness. The NDC diagnostic stage identified high-similarity action pairs, including forestry–coastal ecosystem restoration (0.85), industry–waste circular economy (0.82), power–industry renewable energy (0.78), and water–agriculture efficiency (0.75). The validated keyword lexicon achieved a retrieval precision of 0.83 and recall of 0.79, with mitigation keywords reaching 0.87 precision and adaptation terms scoring 0.76. The study concludes that the Sri Lanka pilot demonstrates the framework's transferability to other nations, transforming noisy sustainability descriptions into verified, retrieval-ready evidence that strengthens HEI contributions to national climate reporting systems.

Keywords: Higher Education Climate Action; Nationally Determined Contributions; Climate Action Keyword Mapping; Climate Policy Text Normalisation; Verified Climate Action Retrieval

INTRODUCTION

Higher Education Institutions (HEIs) are identified and recognised as critical actors in climate governance, with influence extending beyond teaching and research into university operations, sustainability reporting, community engagement, and behavioural transformation. Universities directly contribute to greenhouse gas emissions through energy consumption, transport systems, procurement, waste management, and land-use

practices, while simultaneously shaping long-term climate action through education, research, and professional development. This dual role positions HEIs as both operational emitters and knowledge-based contributors to climate transformation (Findler *et al*., 2019; Tilbury, 2011; UNESCO, 2020).

Under the Paris Agreement, countries communicate mitigation and adaptation commitments through Nationally Determined Contributions (NDCs), which serve as the principal instruments for national climate action planning, reporting, and accountability (UNFCCC, 2015). However, effective NDC implementation extends beyond national ministries to subnational and institutional actors capable of translating broad policy commitments into measurable local actions. In this context, HEIs can meaningfully contribute to national climate objectives through renewable energy adoption, energy efficiency, sustainable transport, waste reduction, water conservation, biodiversity restoration, climate education, and community-based adaptation. When properly documented and aggregated through subnational reporting structures, these institutional actions can strengthen Locally Determined Contributions (LDCs) and enhance national climate reporting.

Sustainability reporting and ranking systems have become important disclosure mechanisms for universities. Platforms such as UI GreenMetric, AASHE STARS, and similar sustainability assessment systems encourage HEIs to report initiatives related to energy, water, waste, transport, education, and governance (AASHE, 2023; UI GreenMetric, 2024). While these systems have improved institutional awareness and sustainability performance, the literature consistently highlights that university sustainability reporting remains weakly verified and insufficiently aligned with national climate policy frameworks (Boiocchi et al., 2023). Consequently, information prepared for sustainability reports or UI GreenMetric submissions is rarely suitable for NDC alignment, carbon quantification, or Measurement, Reporting, and Verification (MRV)-compatible national reporting without further processing.

A fundamental driver of this gap is terminology inconsistency. University climate actions are frequently described using incomplete, broad, or non-standard keywords. For instance, a solar photovoltaic installation may variously be labelled an "energy saving project," a "green campus initiative," a "renewable infrastructure improvement," or a "carbon reduction activity." Although these descriptions may refer to the same mitigation action, their inconsistent wording obstructs automated keyword detection, semantic retrieval, and NDC mapping. Similarly, adaptation-related actions, including rainwater harvesting, mangrove restoration, climate-resilient agriculture, and disaster risk reduction, are often published without explicitly invoking the term "climate adaptation," causing automated retrieval systems and AI-based classifiers to overlook or misclassify them.

This challenge is further compounded by structural limitations within NDC documents themselves. Because NDCs are typically prepared through sector-specific expert processes, they frequently contain structurally similar actions across different sectors and, in some cases, contradictory or competing actions arising from sectoral trade-offs. Renewable energy expansion may overlap with industrial decarbonisation; water efficiency may appear in both agriculture and water-sector adaptation; ecosystem restoration may span forestry, biodiversity, and coastal resilience; and circular economy principles may emerge in both industry and waste-sector mitigation. These overlaps are not inherently erroneous but must be identified and resolved before NDCs are used as a reference framework for automated climate-action mapping.

Two specific structural issues, therefore, arise when aligning HEI climate actions with NDCs. The first is sectoral similarity, where different NDC sectors contain overlapping or near-duplicate action descriptions, risking duplicate mapping, ambiguous classification, or inflated reporting if the same university activity is linked to multiple sectors without clear disambiguation rules. The second issue is action-level conflict, where competing land-use, infrastructure, or resource priorities across NDC sectors create contradictory policy signals. Examples include land-intensive renewable energy development conflicting with forest conservation, coastal tourism infrastructure competing with mangrove restoration, and urban expansion undermining biodiversity protection. If undetected, such conflicts expose automated mapping systems to the risk of generating misleading or overly positive NDC alignments.

Recent advances in Natural Language Processing (NLP), Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG) offer promising pathways for extracting and interpreting weakly published

sustainability information. RAG systems improve factual grounding by retrieving relevant source passages before generating outputs, reducing dependence on model parameters and enabling stronger evidence traceability (Izacard & Grave, 2021; Lewis *et al*., 2020). This is particularly valuable in climate governance, where NDC mapping, policy interpretation, and carbon-related classification demand source-based evidence rather than generative reasoning alone.

This study responds to these challenges by establishing data cleaning and local keyword-set development as foundational requirements for reliable higher education climate-action mapping. The proposed framework processes NDC documents and university sustainability-reporting text, including institutional websites, sustainability reports, and UI GreenMetric-style submissions by removing duplicate or near-duplicate content, normalising terminology, preserving metadata, and constructing a locally relevant keyword set covering mitigation, adaptation, and cross-cutting climate actions. The keyword set is informed by IPCC reporting principles, UNFCCC climate governance structures, NDC sectoral frameworks, SDG 13, UI GreenMetric indicators, and established university sustainability reporting practices (IPCC, 2023; Ministry of Environment, Sri Lanka, 2021; UI GreenMetric, 2024; UNFCCC, 2015; United Nations, 2015).

Sri Lanka is used as the pilot study context because its updated NDCs provide a structured national reference framework for testing sectoral similarity detection, action-level conflict identification, and higher education climate-action keyword mapping (Ministry of Environment, Sri Lanka, 2021). Sri Lanka also represents a relevant developing-country context where climate reporting infrastructure, institutional sustainability reporting capacity, and cross-sectoral policy coordination remain emergent, making it an appropriate environment for developing transferable methodological solutions. By validating the framework in this context, the study demonstrates how noisy, inconsistent sustainability descriptions can be transformed into clearer, retrieval-ready evidence that supports more reliable automated detection, NDC mapping, SDG alignment, and future carbon quantification across comparable national settings.

Research Aim

This study aims to develop and validate a local data cleaning and keyword mapping framework that transforms inconsistent and weakly standardised higher education sustainability descriptions into verified, retrieval-ready evidence for accurate alignment with NDCs, LDCs, SDGs, UI GreenMetric indicators, and future MRV-compatible national climate reporting systems, using Sri Lanka as a pilot study context.

Research Objectives

1. To analyse Sri Lanka's updated NDCs as the national reference framework, applying NLP-based similarity and conflict detection methods to identify structurally overlapping and potentially contradictory climate actions that may compromise the accurate alignment of higher education sustainability activities with national climate commitments.
2. To develop a local data cleaning and keyword mapping framework for processing NDC documents, university sustainability reports, institutional web content, and UI GreenMetric-style evidence, incorporating duplicate removal, terminology normalisation, metadata preservation, and the construction of locally relevant mitigation, adaptation, and cross-cutting climate-action keywords.
3. To evaluate the effectiveness and transferability of the proposed framework using retrieval precision, recall, and category-wise verification metrics, assessing its capacity to improve automated climate-action detection and NDC/SDG alignment for future MRV-compatible reporting.

Research Gap

Although HEIs increasingly disclose sustainability activities through institutional websites, sustainability reports, annual reports, and UI GreenMetric-style submissions, these descriptions frequently remain terminologically inconsistent and poorly aligned with national climate-policy frameworks. Concurrently, NDC documents themselves present structural challenges, including similar actions distributed across multiple sectors and contradictory actions arising from sectoral trade-offs that render direct automated alignment unreliable. Despite

growing interest in AI-assisted climate-action retrieval and sustainability reporting, existing approaches do not sufficiently address the upstream requirement to clean both NDC reference documents and university sustainability text, resolve duplication, preserve metadata, normalise terminology, and maintain a locally relevant climate-action keyword set before automated detection, NDC mapping, SDG alignment, or carbon quantification can be reliably performed. This study, therefore, addresses the critical absence of a structured, locally grounded data cleaning and keyword mapping framework for transforming higher education climate-action descriptions into verified, retrieval-ready evidence for national and subnational climate reporting systems.

RELATED WORK

The intersection of climate governance, institutional sustainability reporting, and automated text analysis has grown considerably as a field of inquiry, yet critical gaps remain in how these streams connect at the level of data preparation and local keyword mapping. This section reviews four interrelated bodies of literature: sustainability reporting in higher education, NDC structure and policy alignment, text similarity and conflict detection in climate-policy documents, and NLP-based retrieval approaches for climate-action mapping.

2.1 Sustainability Reporting in Higher Education

University sustainability reporting has evolved from voluntary disclosure into a structured institutional practice driven by ranking systems, stakeholder expectations, and policy accountability. Ranking frameworks such as UI GreenMetric and AASHE STARS provide indicator-based structures covering energy, water, waste, transport, education, research, and governance, encouraging universities to document and improve sustainability performance (AASHE, 2023; UI GreenMetric, 2024). Boiocchi et al. (2023) noted that these systems have successfully promoted institutional competition and awareness. However, they were not designed as national climate accounting tools, and their indicators do not consistently provide the methodological specificity required for NDC alignment, carbon quantification, or MRV-compatible reporting.

Lozano (2011) identified persistent unevenness in university sustainability reporting, with significant variation in reporting boundaries, indicator selection, data availability, and methodological consistency. Ceulemans *et al*. (2015) reinforced this finding, observing that reports often emphasise descriptive activities rather than independently verified outcomes, limiting their utility for policy alignment. A university may report renewable energy installations, tree planting, waste reduction initiatives, or awareness programmes, yet these activities may not be expressed in a form that automated systems can directly map to NDC sectors, SDG targets, or MRV frameworks. This gap between institutional disclosure and national climate policy recognition represents a core motivating problem for this study.

2.2 NDC Structure, Policy Alignment, and Sectoral Trade-offs

Under the Paris Agreement, NDCs serve as the primary mechanism through which countries communicate national climate commitments across mitigation and adaptation (UNFCCC, 2015). SDG 13 further reinforces the need to embed climate action within national planning, strategies, and institutional practice (United Nations, 2015). However, connecting institutional sustainability reporting to NDC frameworks is not straightforward. NDCs are typically prepared through sector-specific processes involving different technical groups, ministries, and expert committees, producing documents where similar climate concepts may appear across multiple sectors using different wording, and where some actions may implicitly trade off against others.

2.3 Text Similarity and Conflict Detection in Policy Documents

Text similarity techniques offer practical tools for resolving these structural challenges in NDC documents. Jaccard similarity measures lexical overlap between keyword sets and is particularly useful for identifying explicitly shared terminology across policy actions (Jaccard, 1912). Cosine similarity, applied through vector-space models, supports document-level comparison by measuring the angular distance between vectorised text representations, capturing semantic proximity even when similar policy intentions are expressed through different terminology (Salton & Buckley, 1988). For instance, "improve water-use efficiency" and "promote

efficient irrigation systems" may not be lexically identical but may represent closely related adaptation actions. Combining these two approaches enables the detection of both direct lexical overlap and latent semantic similarity, providing a more robust basis for NDC diagnostic analysis.

TF-IDF vectorisation further supports this analysis by weighting terms according to their discriminative value across a document corpus, reducing the influence of generic vocabulary and improving the precision of similarity calculations (Salton & Buckley, 1988). Applied together, these techniques enable a systematic pre-processing stage in which similar and potentially conflicting NDC actions are identified and flagged before they are used as a reference framework for automated HEI climate-action mapping, a step that the existing literature has not formally integrated into sustainability reporting workflows.

2.4 NLP, RAG, and the Data Quality Problem

Recent advances in NLP, Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG) have created new opportunities for analysing the inconsistencies of sustainability and climate-policy texts. RAG systems improve factual grounding by retrieving relevant evidence from a document corpus before generating responses, reducing reliance on the model's internal parametric knowledge (Lewis *et al*., 2020). Izacard and Grave (2021) demonstrated that retrieval-supported generation substantially improves performance on knowledge-intensive tasks by anchoring outputs in externally retrieved passages, a property that is especially important for climate-action mapping, where classification decisions must be traceable to verifiable source evidence.

However, RAG does not automatically guarantee output quality. Ji *et al*. (2023) describe hallucination as the generation of plausible but factually unsupported content and identify it as one of the most significant challenges in deployed LLM systems. In climate governance contexts, hallucination is particularly consequential because incorrect outputs may produce wrong NDC mappings, inaccurate SDG alignments, or misleading assumptions for carbon quantification. Critically, hallucination risk is not solely a model-level problem; it is also shaped by the quality of the retrieved corpus. A document corpus containing duplicated passages, inconsistent terminology, weak metadata, or unresolved policy contradictions increases the probability of poor retrieval, which in turn increases the probability of incorrect generation. Reliable AI-assisted climate-action mapping, therefore, requires upstream data cleaning, controlled vocabulary development, provenance preservation, and local keyword governance before RAG is applied, steps that existing frameworks have not systematically addressed.

A further gap concerns local terminology. Globally standardised climate vocabulary, including terms such as "renewable energy," "carbon neutrality," and "climate resilience", is necessary but insufficient for detecting climate actions in specific national or institutional contexts. In Sri Lankan higher education reporting, climate-relevant activities may be described through locally embedded phrases such as "rainwater harvesting," "solar rooftop," "biofertiliser," "biodiversity garden," "waste segregation," "mangrove restoration," and "sustainable commuting." Without a locally maintained keyword set, automated systems risk missing these actions or classifying them too broadly, undermining both retrieval precision and NDC alignment accuracy.

Collectively, the literature establishes the importance of HEIs in climate governance, the limitations of existing sustainability reporting frameworks, and the potential of NLP and RAG for evidence-based climate analysis. However, no existing approach adequately addresses the upstream requirement to simultaneously clean NDC reference documents, resolve sectoral overlaps and conflicts, normalise university sustainability text, and construct a locally grounded climate-action keyword set.

METHODOLOGY

3.1 Methodological Overview

This study develops a local data cleaning framework for higher education climate-action keyword mapping, transforming inconsistent sustainability descriptions into verified, retrieval-ready evidence for climate-action detection, NDC alignment, SDG mapping, UI GreenMetric reporting support, and future MRV-compatible

carbon quantification (Gunathilake *et al*., 2025). The central methodological premise is that reliable AI-assisted climate-action retrieval cannot begin at the model level; it must begin upstream, with the preparation of a clean, locally grounded, and policy-aligned textual corpus. A noisy or structurally unresolved input corpus undermines retrieval quality regardless of model sophistication.

The framework is operationalised through six sequential stages:

1. Selection of reference documents and institutional evidence sources,
2. Document extraction, cleaning, and normalisation,
3. NDC diagnostic analysis using similarity and conflict detection,
4. Construction of a locally relevant climate-action keyword set,
5. Keyword validation using retrieval performance measures,
6. Preparation of verified, retrieval-ready evidence for automated climate-action mapping.

Sri Lanka has been adopted as the pilot study context because its updated NDCs provide a nationally recognised and structured reference framework covering mitigation, adaptation, and related climate actions (Ministry of Environment, Sri Lanka, 2021). The higher education context is selected because universities increasingly publish sustainability evidence through institutional websites, annual reports, sustainability reports, and UI GreenMetric-style submissions, yet these descriptions frequently lack the terminological consistency required for automated climate-action detection and reliable NDC alignment.

3.2 Data Sources and Corpus Construction

The corpus was constructed from two complementary document categories. The first comprised national climate-policy reference documents, principally Sri Lanka's updated NDCs, which provide the official sectoral structure against which higher education climate actions are interpreted and mapped. The second comprised higher education sustainability evidence, including institutional web content, sustainability reports, annual reports, climate-related project descriptions, and UI GreenMetric-style reporting submissions from Sri Lankan universities.

The NDC corpus was used to identify official sectoral categories, structurally overlapping policy actions, and potentially contradictory actions arising from cross-sectoral trade-offs. The higher education corpus was used to understand how universities describe climate-relevant activities in practice and to identify the terminological gap between institutional reporting language and national climate-policy vocabulary. Because the objective of this study is keyword mapping rather than final carbon quantification, emphasis was placed on textual evidence quality, terminology consistency, and retrieval readiness rather than activity-level quantitative data.

Each document was registered with minimum metadata fields before processing, including document title, issuing institution, document type, publication year, source location, sector classification where available, language, page or section reference, and extraction date. Metadata preservation is not merely administrative; it is a functional requirement for retrieval-based systems where evidence must remain traceable to its source, particularly when AI-assisted interpretation is applied to policy-sensitive climate reporting contexts (Lewis *et al*., 2020; Ji *et al*., 2023).

3.3 Document Extraction, Cleaning, and Normalisation

The first processing stage converted heterogeneous source documents into clean, machine-readable text. Documents were extracted from PDF files, HTML pages, web-based institutional reports, and text-based publications. HTML content was parsed to remove navigation menus, sidebars, repeated headers, footers, and non-substantive boilerplate text. PDF content was processed with page reference preservation, as page-level traceability is essential for downstream verification.

3.3.1 Encoding and whitespace normalisation corrected broken characters, line-break errors, spacing inconsistencies, and Unicode irregularities introduced during PDF-to-text conversion. This was particularly critical for climate-action descriptions split across page boundaries or formatted in multi-column layouts.

3.3.2 Boilerplate removal eliminated repeated institutional slogans, navigation structures, page headers, and non-substantive website text. Without this step, retrieval systems may incorrectly assign high relevance to repeated but informationally empty phrases, distorting evidence ranking.

3.3.3 Terminology normalisation harmonised common variants of climate-related terms to support consistent downstream retrieval. For example, "PV," "solar PV," "photovoltaic," and "solar photovoltaic system" were treated as terminological equivalents. Similarly, "GHG," "greenhouse gas," and "carbon emission" were normalised without discarding the original wording, which was retained in the evidence record for traceability. Table 1 provides illustrative examples of terminology normalisation applied in this framework.

3.3.4 Unit and date normalisation standardised measurement expressions, including energy units (kWh, MWh), capacity (MW), mass (tonnes, kg), area (hectares), distance (km), and emission equivalents (tCO_2e). Consistent unit representation is a prerequisite for future carbon quantification workflows, even though quantitative carbon accounting is beyond the direct scope of this study.

3.3.5 Metadata attachment associated source information with each cleaned text segment. Every extracted segment was stored with a unique identifier, source document reference, page or section number, institutional origin, and extraction date, forming the provenance layer of the retrieval-ready corpus.

3.4 Duplicate and Near-Duplicate Detection

Duplicate and near-duplicate text represents a systematic source of error in automated retrieval systems. The same climate action may appear across multiple institutional reports, website pages, press releases, and annual summaries. When such repetitions are indexed independently, retrieval systems may overrepresent a single activity, creating a misleading impression of breadth or scale of climate engagement. Duplicate prevention was therefore treated as a foundational rather than an optional cleaning step.

Let the fully cleaned corpus be represented as:

$$D = \{d_1, d_2, d_3, \dots, d_n\}$$

where $D$is the document corpus and d_i is the i-th cleaned text segment.

Exact duplicate detection was performed using text-signature comparison on cleaned passages. Near-duplicate detection was applied using cosine similarity between vectorised text segments. Two segments were flagged as near-duplicates when:

$$\text{NearDuplicate}(d_i, d_j) = \begin{cases} 1, & \text{if CosSim}(d_i, d_j) \geq \delta \\ 0, & \text{otherwise} \end{cases}$$

where δ is the near-duplicate similarity threshold. A conservative, high threshold was applied because the objective is not to merge conceptually related actions, which may legitimately describe distinct activities, but to prevent substantially identical text from being indexed as independent evidence.

3.5 NDC Diagnostic Analysis

Before higher education climate actions can be mapped to NDCs, the NDC reference document itself must be diagnostically examined. This step is necessary because NDC documents frequently contain structurally similar

actions distributed across different sectors, and may include potentially contradictory actions arising from competing land-use, resource, infrastructure, or policy priorities. These characteristics are expected in national climate documents prepared through sector-specific expert processes with limited cross-sectoral coordination. The diagnostic stage does not reject or revise the NDC document; rather, it identifies, flags, and minimises the mapping ambiguity that these structural features introduce.

Each NDC action description was treated as a discrete textual unit for analysis. Standard NLP preprocessing was applied in sequence: tokenisation segmented action text into meaningful units; stop-word removal eliminated low-information function words; stemming or lemmatisation reduced related word forms to a common base, improving consistency across action descriptions; and TF-IDF vectorisation transformed preprocessed text into numerical representations suitable for pairwise similarity comparison.

3.6 Jaccard Similarity for Keyword Overlap Detection

Jaccard similarity was applied to identify explicit keyword overlap between NDC sectors. Where K_A and K_B representing the keyword sets extracted from sectors A and B, respectively, the Jaccard similarity coefficient is defined as:

$$J(A,B) = \frac{| K_A \cap K_B |}{| K_A \cup K_B |}$$

where $| K_A \cap K_B |$ is the count of shared keywords and $| K_A \cup K_B |$ is the total count of unique keywords across both sectors (Jaccard, 1912). Values approaching 1 indicate high lexical overlap; values approaching 0 indicate minimal shared vocabulary.

In the Sri Lanka NDC context, Jaccard similarity was applied to detect cases where climate concepts such as renewable energy, circular economy principles, ecosystem restoration, water-use efficiency, and capacity building appeared across multiple sectors using overlapping terminology. High Jaccard scores flagged sector pairs requiring further examination before the NDC document was used as a mapping reference.

3.7 TF-IDF Vectorisation

Term Frequency–Inverse Document Frequency (TF-IDF) was applied to capture the discriminative value of terms within the NDC and university sustainability corpus. TF-IDF reduces the influence of generic terms such as "climate," "sustainable," or "development" that appear frequently across all documents while increasing the weight of terms that are more specific and informative within the corpus.

For a term $t$in document d:

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) \times \log\left(\frac{N}{\text{DF}(t)}\right)$$

where $\text{TF}(t,d)$ is the frequency of term t in document d, N is the total number of documents in the corpus, and $\text{DF}(t)$ is the number of documents containing term t(Salton & Buckley, 1988).

TF-IDF vectorisation was applied both to NDC diagnostic similarity analysis and to the discriminative validation of candidate climate-action keywords. Terms that appeared frequently in climate-relevant institutional evidence but infrequently in general sustainability content were prioritised for inclusion in the local keyword set.

3.8 Cosine Similarity for Semantic Proximity

While Jaccard similarity identifies direct lexical overlap, cosine similarity was employed to detect semantically related NDC actions that may express equivalent policy intentions through different vocabulary. Once action

descriptions were represented as TF-IDF vectors, the cosine similarity between two descriptions d_i and d_j was calculated as:

$$\text{CosSim}(d_i, d_j) = \frac{\overrightarrow{v_i} \cdot \overrightarrow{v_j}}{\| \overrightarrow{v_i} \| \| \overrightarrow{v_j} \|}$$

where $\overrightarrow{v_i}$ and $\overrightarrow{v_j}$ are the TF-IDF vector representations, $\overrightarrow{v_i} \cdot \overrightarrow{v_j}$ is their dot product, and $\| \overrightarrow{v_i} \|$ and $\| \overrightarrow{v_j} \|$ are their respective Euclidean norms (Salton & Buckley, 1988).

Values close to 1 indicate strong semantic similarity; values close to 0 indicate weak or no semantic relationship. This technique was particularly effective for detecting cases where actions such as "improve water-use efficiency" and "promote efficient irrigation systems" share policy intent without employing identical keywords, a distinction that Jaccard similarity alone would not capture.

3.9 Sectoral Similarity Index

To summarise and compare similarity at the sector level, pairwise action-level cosine similarities were aggregated into a Sectoral Similarity Index (SSI)**.** Let S_i and S_j represent two NDC sectors, and let A_i and A_j represent the sets of action descriptions within those sectors. The SSI is defined as:

$$\text{SSI}(S_i, S_j) = \frac{1}{| A_i || A_j |} \sum_{a \in A_i} \sum_{b \in A_j} \text{CosSim}(a, b)$$

The SSI produces a normalised, sector-level similarity score that enables systematic comparison across all NDC sector pairs. Sector pairs exceeding a predefined SSI threshold were flagged as high-overlap sectors and prioritised for closer inspection before being used in the HEI climate-action mapping reference. This step operationalises the transition from action-level diagnostic analysis to sector-level corpus governance.

3.10 NDC Conflict Detection

Semantic similarity between NDC actions does not imply policy coherence. Structurally similar actions may reinforce each other, but they may equally compete for land, resources, finance, or regulatory priority. A dedicated conflict detection procedure was therefore applied to distinguish reinforcing from opposing NDC actions.

Conflict detection combined semantic similarity with policy polarity analysis. Policy polarity vectors $P(a)$and $P(b)$were derived for each action using directional policy terms, including *expand*, *restrict*, *conserve*, *restore*, *reduce*, *intensify*, *protect*, *promote*, and *develop* that indicate the intended direction of a policy action. A conflict was flagged when two actions were semantically proximate but expressed opposing policy directions:

$$\text{Conflict}(a, b) = \begin{cases} 1, & \text{if CosSim}(a, b) \geq \theta \ \wedge \ P(a) \cdot P(b) < 0 \\ 0, & \text{otherwise} \end{cases}$$

where θ is the semantic similarity threshold and $P(a) \cdot P(b) < 0$ denotes opposing polarity. Illustrative conflict pairs in the Sri Lanka NDC context include land-intensive renewable energy expansion versus forest conservation, coastal infrastructure development versus mangrove restoration, and irrigation expansion versus water-resource sustainability commitments. Flagging such conflicts prevents automated systems from treating apparent keyword similarity as policy alignment, which would otherwise generate misleading NDC mappings for university activities.

3.11 Local Climate-Action Keyword Set Construction

A locally grounded keyword set was constructed to support the automated detection of higher education climate actions within Sri Lankan institutional reporting. The keyword set was organised into three primary classification categories: mitigation**,** adaptation**,** and cross-cutting climate action, consistent with the structural conventions of NDC reporting and international climate governance (UNFCCC, 2015; Ministry of Environment, Sri Lanka, 2021).

3.11.1 Seed keyword extraction drew from official and policy-relevant sources, including Sri Lanka's NDC terminology, SDG 13 language, UI GreenMetric indicators, and established university sustainability reporting categories. This ensured the keyword set reflected both national climate priorities and higher education sustainability reporting conventions (United Nations, 2015; UI GreenMetric, 2024).

3.11.2 Local terminology enrichment extended the seed set to capture climate-action phrases characteristic of Sri Lankan university reporting contexts. These included terms such as solar rooftop*,* rainwater harvesting*,* biodiversity garden*,* waste segregation*,* mangrove restoration*,* organic farming*,* biofertiliser*,* green procurement*,* sustainable commuting*,* and community engagement phrases that globally standardised climate vocabularies would not reliably capture.

3.11.3 Semantic expansion identified related phrases, terminological variants, and contextually equivalent expressions. However, expanded terms were not automatically admitted to the keyword set, as broad sustainability vocabulary carries a significant risk of increasing false-positive retrieval rates.

3.11.4 Empirical and manual filtering were applied as the final admission criterion. A term was retained when it demonstrably improved climate-action detection precision and could be unambiguously mapped to at least one mitigation, adaptation, or cross-cutting category. Contextually ambiguous terms such as *green*, *awareness*, or *resilience* were classified as context-restricted keywords, meaning they were activated for retrieval only when co-occurring with stronger, unambiguous climate-action terms.

3.12 Keyword Classification Structure

Each keyword was stored with structured metadata to enable controlled, category-aware retrieval. The keyword record schema is presented in Table 1. This schema enables the keyword set to function as a governed, maintainable vocabulary rather than a static list. Maintaining keyword governance over time is essential because climate-action terminology evolves with successive NDC revisions, updated IPCC reporting guidelines, and changes in institutional sustainability reporting practices.

Table 1: Keywords' metadata fields

Field	Description
keyword_id	Unique identifier for the keyword entry
keyword	Climate-action keyword or phrase
category	Mitigation, adaptation, or cross-cutting
sub-theme	Energy, transport, water, waste, biodiversity, education, governance, etc.
source_basis	NDC, SDG, UI GreenMetric, IPCC/UNFCCC terminology, or local reporting usage
Retrieval_status	Active, context-restricted, under review, or deprecated
mapping_notes	Relevance notes for NDC, SDG, or UI GreenMetric alignment

3.13 Retrieval-Ready Evidence Preparation

Following cleaning, duplicate detection, NDC diagnostic analysis, conflict flagging, and keyword validation, the final output of the framework is a retrieval-ready evidence corpus. Each evidence chunk is stored with the

structured fields presented in Table 2.

Table 2: NDC duplicate action detection record storing schema

Field	Description
chunk_id	Unique evidence identifier
cleaned_text	Cleaned and normalised text passage
source_document	Original document or institutional web source
page_or_section	Page number, section heading, or URL reference
institution	University or issuing body
keyword_matches	Matched local climate-action keywords
climate_category	Mitigation, adaptation, or cross-cutting
ndc_sector	Relevant NDC sector where applicable
sdg_mapping	Related SDG category, where applicable
similarity_links	Related NDC actions or sectors identified in diagnostic analysis
conflict_flags	Potential action-level conflicts identified during the NDC diagnostic stage
duplicate_group	Duplicate or near-duplicate group identifier
retrieval_weight	Evidence priority score for RAG or retrieval system use

The retrieval-ready corpus is designed for direct integration into RAG pipelines or other AI-assisted retrieval systems. Critically, retrieval is constrained by cleaned evidence, local keyword governance, and full source traceability, reducing the probability that AI systems generate unsupported, ambiguously mapped, or hallucinated climate-action classifications.

RESULTS AND DISCUSSION

4.1 Overview of Results and Discussion

The results of this study are presented in accordance with the six-stage methodological workflow developed for higher education climate-action keyword mapping, using Sri Lanka's updated NDCs as the national reference framework and Sri Lankan university sustainability descriptions as the institutional evidence corpus. The findings address the central research problem from two complementary directions: the structural quality of the national climate-policy reference framework, and the terminological quality of sustainability descriptions published by higher education institutions. The results demonstrate that weaknesses in either layer whether manifesting as sectoral overlaps, policy contradictions, duplicated passages, inconsistent terminology, or locally absent vocabulary, propagate directly into automated retrieval systems, producing incorrect NDC alignment, weak SDG mapping, and unreliable evidence for future MRV-compatible carbon quantification. This confirms that upstream data preparation is not a peripheral preprocessing activity but a methodologically indispensable foundation for any AI-assisted climate-action mapping pipeline.

The Sri Lanka pilot generated five principal and interconnected outcomes. First, the NDC diagnostic analysis revealed structurally similar climate actions distributed across multiple sectors, establishing that NDC reference documents carry inherent mapping ambiguity that must be resolved before they can serve as reliable alignment frameworks for automated systems. Second, the conflict-detection stage identified action-level trade-offs where semantically proximate NDC actions express opposing policy directions, a finding that challenges simplistic keyword-matching approaches and demonstrates the necessity of polarity-aware interpretation. Third, the document cleaning and normalisation process materially improved the structural integrity of both the climate-policy corpus and the university sustainability evidence base, through duplicate removal, terminology harmonisation, and systematic metadata preservation. Fourth, the locally constructed keyword set substantially improved the detectability of university climate actions across mitigation, adaptation, and cross-cutting categories, capturing practice-based and regionally specific terminology that globally standardised climate vocabularies consistently fail to retrieve. Fifth, the retrieval validation stage confirmed that the keyword lexicon achieved an overall precision of 0.83 and recall of 0.79, with category-wise performance diverging between

mitigation and adaptation terms in ways that carry direct implications for keyword governance and retrieval system design.

4.2 Outcome of Reference Document and Institutional Evidence Selection

The first stage of the methodology established the documentary foundation for climate-action keyword mapping by selecting and registering the national climate-policy reference documents and higher education institutional evidence sources against which the framework was developed and validated.

Sri Lanka's updated NDCs were adopted as the national reference framework on the basis that NDCs constitute the formal policy instrument through which countries communicate quantified mitigation commitments and structured adaptation priorities under the Paris Agreement (UNFCCC, 2015; Ministry of Environment, Sri Lanka, 2021). As an official, internationally submitted document, the Sri Lanka NDC provides a sector-structured, action-level vocabulary that represents the most authoritative available reference for determining whether a university activity constitutes a nationally recognised climate action. The NDC corpus was supplemented with SDG 13 documentation and UI GreenMetric indicator frameworks to ensure that the reference structure captured both global sustainability reporting conventions and national climate-policy priorities.

The higher education evidence corpus was assembled from multiple institutional source types, including university websites, sustainability reports, annual reports, project and programme descriptions, environmental and sustainability policy documents, and UI GreenMetric-style reporting submissions. This multi-source approach was necessary because Sri Lankan universities weakly publish climate-relevant activities through a single standardised reporting channel. Climate actions are instead distributed across heterogeneous document formats, institutional web structures, and reporting cycles that vary significantly across institutions.

This distributional inconsistency was itself a significant finding of the selection stage. A renewable energy installation, for instance, may be disclosed as a news announcement on an institutional website, as a line item in a sustainability report, as supporting evidence in a UI GreenMetric submission, or as a procurement notice, with each instance using different descriptive language, varying levels of technical detail, and inconsistent or absent keyword specificity. Adaptation-related activities presented an even greater challenge: actions such as rainwater harvesting systems, biodiversity gardens, watershed protection programmes, and community health preparedness initiatives were frequently described under headings related to community service, environmental conservation, or institutional development, without any explicit reference to climate adaptation as a policy category. This pattern confirms that the terminological gap between how universities describe their activities and how national climate frameworks categorise them is not incidental but structurally embedded in current institutional reporting practices.

The metadata registration process, applied to every source document before further processing, revealed additional inconsistencies in document completeness. Publication dates were absent or ambiguous in several institutional sources, sector classifications were not uniformly applied, and source traceability essential for downstream retrieval verification was inconsistently maintained across document types. These observations reinforced the methodological decision to treat metadata preservation as a functional requirement of the cleaning framework rather than an administrative supplement. Without reliable provenance information attached to each evidence segment, AI-assisted retrieval systems would face problems in distinguishing between current and outdated institutional commitments, or between verified activities and aspirational policy statements, a distinction that carries direct implications for the credibility of NDC alignment and future MRV-compatible reporting.

Table 3: Selected evidence sources and their relevance to climate action keyword mapping

Evidence source	Main content type	Relevance to keyword mapping	Main limitation
Updated NDC document	National mitigation and adaptation actions	Provides official climate-policy reference categories	Contains similar or conflicting actions across sectors

NDA or climate-policy documents	Governance and implementation guidance	Supports institutional and national alignment	Terminology may differ from the university reporting language
University websites	News, project pages, policy pages, announcements	Provides discoverable evidence for web crawlers	Often incomplete, inconsistent, or unstructured
Sustainability reports	Institutional sustainability achievements	Provides annual or periodic climate-action descriptions	Descriptive rather than quantification-ready
UI GreenMetric-style evidence	Ranking-related sustainability indicators	Supports structured reporting across energy, water, waste, transport, education, and governance	Not directly designed for NDC or MRV alignment
Annual reports	Institutional projects and development activities	Includes infrastructure, energy, and environmental actions	Climate relevance may not be explicitly stated

As shown in Table 3, the results indicate that higher education climate evidence already exists, but it is not automatically usable for national climate reporting. The main challenge is not the absence of activities, but the lack of structured, standardised, and locally aligned descriptions. This supports the study's argument that a data cleaning and keyword mapping framework is required before automated detection, NDC alignment, or future carbon quantification can be performed.

4.3 Results of Document Extraction, Cleaning, and Normalisation

The second methodological stage processed the assembled corpus of national climate-policy documents and university sustainability evidence through systematic extraction, cleaning, and normalisation procedures. This stage addressed the full range of textual quality issues identified across heterogeneous source types, including repeated headers and footers, broken character sequences introduced during PDF-to-text conversion, duplicated paragraphs appearing across multiple institutional publications, inconsistent spelling and capitalisation of climate-related terms, mixed and non-standard measurement expressions, and incomplete or absent metadata fields. Left unresolved, each of these issues introduces a distinct failure mode into downstream retrieval and keyword mapping operations. The cleaning and normalisation process produced four substantive and interconnected improvements to corpus quality.

The first improvement was a material reduction in retrieval noise. Non-substantive text, including website navigation menus, repeated institutional footers, campus event announcements, generic sustainability slogans, and boilerplate procurement language, was systematically identified and removed. This content is particularly problematic for automated retrieval systems because it may contain surface-level climate-related terms, such as "green," "sustainable," or "environment," without conveying any substantive climate-action evidence. Without removal, such text inflates the apparent volume of climate-relevant content, increases false-positive retrieval rates, and distorts evidence ranking within RAG or similar retrieval pipelines. The boilerplate removal stage ensured that the corpus retained only passages with genuine informational content relevant to climate-action detection and NDC alignment.

The second improvement was a significant increase in terminology consistency across the corpus. University sustainability documents exhibited considerable variation in how equivalent climate actions were named and described, reflecting differences in institutional reporting conventions, authorship styles, and the absence of a shared controlled vocabulary. Terminological variants referring to the same underlying technology or practice such as "solar PV," "photovoltaic system," "solar rooftop installation," and "PV panels" within the energy category, or "GHG inventory," "greenhouse gas accounting," and "carbon emissions measurement" within the mitigation monitoring category were mapped into coherent keyword families while preserving the source

wording within each evidence record. This dual-layer approach, normalisation for retrieval consistency, preservation for source traceability, is essential in climate governance contexts where the precise language used in institutional disclosures may carry policy or verification significance.

The third improvement concerned the standardisation of quantitative expressions in preparation for future carbon quantification workflows. Activity-level numerical data, including energy generation and consumption figures (kWh, MWh), installed capacity (MW, kW), mass-based waste and emissions data (kg, tonnes, tCO_2e), land area (hectares), and transport distance (kilometres) were normalised to consistent unit formats across all evidence segments. While final carbon accounting lies beyond the direct scope of this study, unit normalisation at the cleaning stage is a prerequisite for any subsequent MRV-compatible quantification process. Inconsistent unit expressions across institutional reports, such as mixing MWh and kWh within the same evidence set, or reporting waste diversion in both weight and volume, would otherwise introduce systematic errors into activity-based emission factor calculations.

The fourth improvement was the systematic preservation of structured metadata for every cleaned evidence segment. Each passage was stored with a unique chunk identifier, source document reference, page or section number, institutional origin, document type, extraction date, and preliminary climate-action category tag. This metadata layer serves a dual function within the framework. Operationally, it enables retrieval systems to return not only relevant evidence passages but also the provenance information required to verify their institutional and documentary origin, a critical requirement for policy-sensitive NDC alignment and MRV reporting contexts. Methodologically, it ensures that the framework remains auditable and reproducible, allowing individual evidence decisions to be traced back to specific source locations rather than treated as outputs of an opaque processing pipeline.

Table 4: Document cleaning outcomes and their relevance to automated climate action mapping

Cleaning function	**Problem addressed**	**Resulting improvement**	**Relevance to automated mapping**
Boilerplate removal	Menus, headers, footers, slogans, repeated web text	Reduced retrieval noise	Improves precision by reducing irrelevant matches
Duplicate removal	Same activity repeated across reports or web pages	Prevented repeated evidence inflation	Reduces double-counting and duplicate NDC mapping
Terminology normalisation	Different words are used for the same action	Improved consistency across documents	Supports more accurate keyword detection
Unit normalisation	Mixed formats for kWh, MW, tonnes, hectares, tCO_2e	Prepared evidence for later quantification	Reduces calculation errors in future MRV use
Metadata preservation	Missing source traceability	Each text segment is linked to the source, page, and institution	Supports evidence verification and RAG grounding
Segment preparation	Long unstructured text blocks	Created retrieval-ready evidence chunks	Improves AI retrieval and classification reliability

Collectively, the outcomes of this stage confirm that document extraction and cleaning cannot be treated as a routine technical preprocessing step in climate-action mapping workflows. The volume and variety of textual quality issues identified across both the national policy corpus and the university evidence base demonstrate that cleaning decisions directly determine what evidence is available to downstream retrieval and classification

systems and therefore directly shape the accuracy and credibility of NDC alignment, SDG mapping, and sustainability reporting outcomes (Table 4).

Further, according to Table 4, cleaning is not merely a technical operation; it is a governance requirement. In climate reporting, an activity becomes useful only when it can be detected, traced, interpreted, and verified. A university may implement a valid climate action, but if the description is vague or the evidence source is not preserved, automated systems may either ignore it or classify it incorrectly. Therefore, text cleaning directly strengthens the reliability of climate action detection.

4.4 Results of NDC Diagnostic Analysis Using Similarity Detection

The third methodological stage analysed the NDC reference framework using text similarity methods. This stage was necessary because NDC documents may contain similar actions across sectors. Such similarity is expected because climate action is cross-sectoral, but it creates ambiguity when automated systems attempt to map university activities to national climate commitments. The analysis used Jaccard similarity to identify explicit keyword overlap and cosine similarity to identify deeper semantic proximity. Jaccard similarity is useful when two policy actions share terms such as “renewable energy,” “ecosystem restoration,” or “water efficiency” (Jaccard, 1912). Cosine similarity, applied through vector-space modelling, supports comparison where similar meanings are expressed using different words (Salton & Buckley, 1988).

The diagnostic stage identified four high-similarity NDC action pairs. The highest similarity score was observed between forestry-sector mitigation and coastal-sector adaptation actions involving ecosystem restoration, especially mangrove restoration, with a similarity score of 0.85. Industry-sector circular economy actions and waste-sector recycling actions showed a similarity score of 0.82. Power-sector renewable energy expansion and industry-sector renewable energy application showed a similarity score of 0.78. Water-sector rainwater harvesting and agriculture-sector irrigation efficiency showed a similarity score of 0.75.

Table 5: High-similarity (75% and above) NDC action pairs identified during diagnostic analysis

Source sector	Related sector	Similar action theme	Similarity score	Interpretation
Forestry mitigation	Coastal adaptation	Forest, mangrove, and ecosystem restoration	0.85	Strong cross-sector overlap between mitigation and adaptation
Industry mitigation	Waste mitigation	Circular economy, recycling, and waste reduction	0.82	Shared resource-efficiency and waste-reduction logic
Power mitigation	Industry mitigation	Renewable energy and solar or wind applications	0.78	Energy transition actions appear across multiple sectors
Water adaptation	Agriculture adaptation	Rainwater harvesting and irrigation efficiency	0.75	Water resilience actions overlap with agricultural adaptation

Table 5 results show that similarity in NDCs should not automatically be treated as a defect. In many cases, overlapping actions reflect genuine interdependencies. For example, mangrove restoration contributes to coastal resilience, biodiversity protection, carbon sequestration, and disaster risk reduction. Similarly, circular economy actions reduce industrial resource consumption while also contributing to waste-sector mitigation. However, from an automated mapping perspective, such overlaps can create duplicate or incorrect alignments. A university tree-planting or mangrove project may be mapped to forestry, coastal adaptation, biodiversity, disaster risk reduction, or SDG 13, depending on keyword choice. If no mapping rule exists, the same activity may be counted multiple times or assigned to an inappropriate primary sector. Therefore, the result supports the need for a

cleaned NDC reference layer in which overlapping sectors are explicitly flagged before automated HEI mapping is performed.

4.5 Results of NDC Conflict Detection

The conflict-detection stage extended the diagnostic analysis beyond similarity. Similarity analysis identifies related actions, but it does not reveal whether those actions are mutually supportive or potentially conflicting. Therefore, conflict detection was used to identify cases where policy actions may compete for land, infrastructure, water, finance, ecosystem space, or implementation priority.

Table 6: Potential NDC action-level conflicts identified for conflict-aware mapping

Climate action theme	Potentially conflicting theme	Conflict type	Mapping implication
Renewable energy expansion	Forest or ecosystem conservation	Land-use competition	Requires site-specific evidence before positive alignment
Electric vehicle promotion	Limited renewable grid capacity	Infrastructure dependency	Transport mitigation depends on energy-sector readiness
Coastal tourism or infrastructure development	Mangrove restoration and coastal resilience	Ecosystem degradation risk	Requires adaptation and biodiversity screening
Irrigation expansion for drought resilience	Water-resource conservation	Resource competition	Requires water balance and context-specific interpretation
Peri-urban development	Forest-cover increase and biodiversity protection	Land-use and biodiversity trade-off	Requires a conflict flag before NDC mapping
Plastic recycling promotion	Single-use plastic reduction	Policy inconsistency risk	Requires distinction between recycling and reduction strategies

The results show (Table 6) that several NDC themes may generate policy trade-offs if interpreted without cross-sectoral review. Renewable energy expansion can support mitigation, but land-intensive renewable energy development may conflict with forest conservation or biodiversity protection. Coastal tourism infrastructure can support economic resilience, but it may conflict with mangrove restoration or coastal ecosystem protection. Irrigation expansion can support agricultural adaptation, but it may increase pressure on water resources. Urban expansion may support settlement needs, but it can create pressure on forest cover and biodiversity.

Further, Table 6 demonstrates that automated climate action mapping should not only identify where a university activity appears to align with an NDC. It should also detect whether the alignment is conditional, partial, or potentially conflicted. For example, a university may construct a renewable energy facility, but if the project description lacks location, land-use, or environmental context, the system should avoid overclaiming its alignment. Similarly, a coastal development activity may be environmentally positive only if it supports ecosystem resilience rather than replacing natural protective systems. This finding strengthens the argument for conflict-aware retrieval. AI-assisted mapping systems should not retrieve only the most similar NDC action; they should also retrieve related conflict flags where relevant. This reduces the risk of overclaiming and supports more responsible climate reporting.

4.6 Results of Local Climate Action Keyword Set Construction

The fourth stage constructed a local higher education climate-action keyword set. The keyword set was organised into three major categories: mitigation, adaptation, and cross-cutting actions. It was populated using terminology

aligned with NDCs, SDG 13, IPCC/UNFCCC climate concepts, UI GreenMetric reporting areas, and locally used university sustainability descriptions. The results show that a locally grounded keyword set is essential because universities often describe climate actions using practical institutional language rather than formal policy language. For example, a university may report "rainwater harvesting," "solar rooftop," "green procurement," "biodiversity garden," or "waste segregation" without explicitly naming the activity as mitigation or adaptation. A generic AI retrieval system may miss these activities or classify them too broadly. The local keyword set bridges this gap by connecting practical university terms to policy-relevant categories.

Table 7: Locally relevant higher education climate-action keyword classes

Keyword class	Sub-theme	Representative keywords and phrases	Expected mapping relevance
Mitigation	Energy and power	renewable energy, solar, photovoltaic, solar rooftop, wind power, LED lighting, smart meter, energy efficiency, carbon footprint	NDC mitigation, SDG 7, SDG 13, UI GreenMetric energy and climate change
Mitigation	Transport	electric vehicle, EV charger, bicycle lane, sustainable commuting, low-emission fleet, public transport	NDC transport mitigation, SDG 11, SDG 13
Mitigation	Waste and circular economy	waste segregation, recycling, composting, circular economy, 3R, food waste reduction, green procurement	NDC waste mitigation, SDG 12, UI GreenMetric waste
Mitigation	Forestry and land use	afforestation, reforestation, tree planting, carbon sink, biodiversity garden, forest restoration	NDC forestry, SDG 15, SDG 13
Adaptation	Water resources	rainwater harvesting, water reuse, leak detection, drought resilience, water security, irrigation efficiency	NDC water adaptation, SDG 6, SDG 13
Adaptation	Agriculture and food security	climate-resilient crops, drought-resistant crops, flood-tolerant crops, organic farming, biofertiliser	NDC agriculture adaptation, SDG 2, SDG 13
Adaptation	Ecosystems and coastal resilience	mangrove restoration, wetlands, coral reef restoration, coastal resilience, ecosystem restoration	NDC coastal/biodiversity adaptation, SDG 14, SDG 15
Adaptation	Health and infrastructure	heat stress, public health adaptation, resilient infrastructure, urban greening, stormwater management	NDC health/urban adaptation, SDG 3, SDG 11
Cross-cutting	Education and governance	climate education, sustainability policy, ESG, green office, MRV, NDC target, Paris Agreement, UNFCCC	SDG 4, SDG 13, UI GreenMetric education and governance
Cross-cutting	Community engagement	awareness programme, climate literacy, student climate action, community engagement, behavioural change	SDG 13, university outreach and reporting

The outcome Table 7 is a structured vocabulary that supports both detection and reporting. It allows automated systems to identify climate-relevant content more accurately, and it gives universities a practical guide for writing better descriptions. The keyword set, therefore, functions not only as a technical retrieval tool but also as a reporting standardisation aid.

4.7 Results of Keyword Validation Using Retrieval Performance Measures

The fifth methodological stage evaluated the keyword set using retrieval performance measures. Precision and recall were used because they directly indicate the performance of keyword-based retrieval. Precision measures how many retrieved passages are actually climate relevant, while recall measures how many climate-relevant passages are successfully retrieved. The validated keyword lexicon achieved an overall precision of 0.83 and recall of 0.79. This indicates that 83% of retrieved passages contained substantive climate-action evidence, while 79% of the relevant climate-action passages were successfully detected (Table 8). The overall F1-score of 0.81 indicates that the keyword set achieved a strong balance between precision and recall. Mitigation keywords performed better than adaptation keywords, with mitigation precision reaching 0.87. This is because mitigation terms such as solar, photovoltaic, EV, LED lighting, carbon footprint, and greenhouse gas are more technically specific. Adaptation keywords achieved a lower precision value of 0.76 because adaptation language overlaps with general environmental, disaster-management, health, agriculture, and infrastructure vocabulary.

Table 8: Keyword validation results

Performance indicator	Result	Interpretation
Overall retrieval precision	0.83	Most retrieved passages contained valid climate-action evidence
Overall retrieval recall	0.79	Most relevant climate-action passages were successfully detected
Overall F1-score	0.81	Balanced retrieval performance
Mitigation keyword precision	0.87	Strong performance due to technically specific terminology
Adaptation keyword precision	0.76	Lower precision due to overlap with general resilience and environmental terms

The keyword validation results in Table 8 reveal that mitigation keyword mapping is more straightforward than adaptation keyword mapping. Mitigation actions usually involve measurable technologies or activities, such as renewable electricity generation, transport electrification, waste recycling, and energy efficiency. Adaptation actions are broader and more context-dependent. Terms such as resilience, preparedness, water security, urban greening, and public health may refer to climate adaptation, but they may also appear in non-climate contexts.

4.8 Outcome of Verified Retrieval-Ready Evidence Preparation

The final methodological stage prepared the cleaned and validated evidence for automated climate-action mapping. Each evidence segment was linked to source metadata, matched keywords, climate category, NDC sector, SDG relevance, duplicate status, and conflict flags. This structure (Table 9) makes the corpus suitable for retrieval-based AI systems while reducing the risk of unsupported outputs.

Table 9: Structure and purpose of the retrieval-ready evidence output

Evidence field	Purpose	Contribution to reliability
Chunk ID	Provides unique evidence of identity	Prevents confusion among similar text segments
Cleaned text	Stores normalised climate-action description	Improves retrieval accuracy
Source document or URL	Links evidence to the source	Enables verification

Page or section reference	Preserves location within source	Supports auditability
Institution	Identifies reporting HEI	Supports institutional aggregation
Keyword matches	Shows detected local climate-action terms	Supports explainable retrieval
Climate category	Mitigation, adaptation, or cross-cutting	Supports classification
NDC sector	Links action to the national climate policy area	Supports NDC alignment
SDG mapping	Links action to development goals	Supports SDG reporting
UI GreenMetric relevance	Links action to reporting category	Supports university sustainability reporting
Duplicate group	Identifies repeated or near-repeated text	Prevents double-counting
Conflict flag	Indicates potential policy trade-off	Supports human review
Retrieval weight	Prioritises stronger evidence	Improves AI retrieval quality

The retrieval-ready evidence structure is important because it converts sustainability descriptions into machine-readable, traceable, and policy-aligned data objects. This does not replace human verification, but it reduces the manual burden by ensuring that AI systems retrieve stronger evidence and that reviewers can trace outputs to sources.

The central finding of this study is that reliable higher education climate-action mapping depends on the quality of both the policy reference framework and the institutional reporting language. NDC documents provide the national structure for climate alignment, but they may contain sectoral overlaps and policy trade-offs. University reports provide institutional evidence, but they may use incomplete or inconsistent terminology. Automated AI systems can connect these two layers only when both are cleaned, normalised, and locally keyword-mapped.

The results also show that correct keyword mapping improves more than retrieval. It improves institutional reporting quality, supports SDG and UI GreenMetric alignment, reduces incorrect NDC mapping, and prepares evidence for future carbon quantification. However, the findings also caution that keywords alone are not enough. A detected keyword must be interpreted in context, supported by source evidence, checked against NDC overlaps, and screened for possible conflict. Therefore, this study contributes a practical and methodological framework for transforming noisy higher education sustainability descriptions into verified, retrieval-ready evidence. It shifts the focus from AI model output to evidence quality. In doing so, it supports a more reliable pathway for integrating HEI climate actions into national climate reporting systems and future MRV-compatible carbon accounting.

CONCLUSIONS

This study developed and validated a local data cleaning framework for higher education climate-action keyword mapping, addressing a critical but underexamined upstream requirement in AI-assisted climate governance: the systematic preparation of both national climate-policy documents and institutional sustainability descriptions before automated retrieval, NDC alignment, SDG mapping, or carbon quantification can be reliably performed. Using Sri Lanka as a pilot study context, the framework demonstrated that data quality rather than model sophistication alone is the foundational determinant of reliable climate-action detection and policy alignment.

Three broader contributions emerge from this work. First, the study establishes data cleaning and local keyword governance as prerequisites rather than supplementary steps in the AI-assisted climate-action mapping pipeline.

Retrieval-Augmented Generation and similar systems cannot compensate for an underlying corpus that contains duplicated passages, unresolved NDC conflicts, inconsistent terminology, or weak metadata provenance. The framework addresses these failure modes systematically and upstream. Second, the Sri Lanka pilot provides an empirically grounded demonstration of the framework's applicability in a developing-country context where climate reporting infrastructure, cross-sectoral coordination, and institutional sustainability reporting capacity remain emergent, precisely the conditions under which structured data preparation offers the greatest marginal benefit. Third, the modular architecture of the framework encompassing document extraction, duplicate detection, NDC diagnostic analysis, keyword construction, and retrieval-ready corpus preparation is explicitly designed for transferability to other national contexts, enabling adaptation to different NDC structures, reporting vocabularies, and HEI sustainability disclosure practices with minimal reconfiguration.

Several directions warrant further investigation. The current framework operates at the keyword and passage level; integration with fine-tuned language models or domain-adapted embeddings could further improve semantic retrieval precision for adaptation-related terminology. Extension of the conflict-detection mechanism to include quantitative polarity scoring, rather than binary flagging, would enable more nuanced prioritisation of ambiguous NDC action pairs. Longitudinal application of the framework across successive NDC revision cycles would also allow assessment of keyword set stability and the extent to which terminological drift in national climate documents affects retrieval performance over time. Finally, scaling the framework beyond Sri Lanka to encompass multiple developing-country NDC structures would provide comparative evidence for its generalisability and support the development of a shared, regionally maintained climate-action keyword resource for higher education sustainability reporting.

Ultimately, this study argues that the integrity of higher education climate reporting and its contribution to national and subnational climate governance depend not only on what universities do, but on how precisely and consistently those actions are described, cleaned, and mapped. The proposed framework provides a replicable, evidence-based pathway for transforming fragmented and terminologically inconsistent sustainability descriptions into verified, retrieval-ready evidence, enabling HEIs to contribute more credibly and more visibly to the climate commitments their countries have made.

FUTURE WORK AND RECOMMENDATIONS

i. Future research should incorporate advanced NLP approaches, including transformer-based and domain-adapted language models, to enhance semantic accuracy and contextual understanding of climate actions. Extensions to multilingual datasets and cross-regional comparative studies are also recommended to improve the framework’s applicability and transferability across diverse institutional and governance environments.
ii. Develop a multilingual keyword set: covering English and other local languages, climate-action terminology to improve detection of HEI sustainability activities published in local languages.
iii. Create a national HEI climate-action keyword registry: with version-controlled keywords classified as active, context-restricted, under review, or deprecated.
iv. Test the framework with RAG-based AI systems: by comparing raw documents, cleaned documents, and locally keyword-mapped documents to measure improvements in retrieval accuracy and factual grounding.
v. Introduce a standard HEI climate-action reporting template: including action type, location, date, responsible unit, activity value, unit of measurement, SDG relevance, NDC sector, and evidence source.

REFERENCES

1. Association for the Advancement of Sustainability in Higher Education. (2023). *STARS technical manual version 2.2*. Association for the Advancement of Sustainability in Higher Education. https://stars.aashe.org/wp-content/uploads/2019/07/STARS-2.2-Technical-Manual.pdf
2. Boiocchi, R., Ragazzi, M., Torretta, V., & Rada, E. C. (2023). Critical analysis of the GreenMetric World University Ranking System: The issue of comparability. *Sustainability, 15*(2), Article 1343. https://doi.org/10.3390/su15021343

3. Ceulemans, K., Molderez, I., & Van Liedekerke, L. (2015). Sustainability reporting in higher education: A comprehensive review of the recent literature and paths for further research. *Journal of Cleaner Production, 106*, 127–143. https://doi.org/10.1016/j.jclepro.2014.09.052
4. Findler, F., Schönherr, N., Lozano, R., Reider, D., & Martinuzzi, A. (2019). The impacts of higher education institutions on sustainable development: A review and conceptualization. *International Journal of Sustainability in Higher Education, 20*(1), 23–38. https://doi.org/10.1108/IJSHE-07-2017-0114
5. Gunathilake, P. M. P. C., Hewawasam, T., & Gunatilake, J. (2025). Aligning and quantifying higher education institutions' climate actions with Nationally Determined Contributions through AI-enabled data discovery and verification. *International Journal of Research and Innovation in Social Science, 9*(26), 9984–9997. https://doi.org/10.47772/IJRISS.2025.903SEDU0766
6. Intergovernmental Panel on Climate Change. (2023). *Climate change 2023: Synthesis report*. IPCC. https://doi.org/10.59327/IPCC/AR6-9789291691647
7. Izacard, G., & Grave, É. (2021). Leveraging passage retrieval with generative models for open domain question answering. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 874–880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.74
8. Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist, 11*(2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x
9. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730
10. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.
11. Lozano, R. (2011). The state of sustainability reporting in universities. *International Journal of Sustainability in Higher Education, 12*(1), 67–78. https://doi.org/10.1108/14676371111098311
12. Ministry of Environment, Sri Lanka. (2021). *Updated Nationally Determined Contributions under the Paris Agreement on climate change: Sri Lanka 2021*. Climate Change Secretariat, Ministry of Environment. https://www.climatechange.lk/CCS2021/NDC%202021%20-%20English.pdf
13. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0
14. Tilbury, D. (2011). *Education for sustainable development: An expert review of processes and learning*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000191442
15. UI GreenMetric World University Rankings. (2024). *UI GreenMetric World University Rankings guideline 2024*. Universitas Indonesia. https://green.rmutk.ac.th/wp-content/uploads/2024/06/UI-GreenMetric-Guideline-2024.pdf
16. United Nations. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development* (A/RES/70/1). United Nations. https://digitallibrary.un.org/record/3923923
17. United Nations Educational, Scientific and Cultural Organization. (2020). *Education for sustainable development: A roadmap*. UNESCO. https://doi.org/10.54675/YFRE1448
18. United Nations Framework Convention on Climate Change. (2015). *Paris Agreement*. United Nations. https://unfccc.int/sites/default/files/english_paris_agreement.pdf