Cleaner Environmental Systems 18 (2025) 100288

Contents lists available at ScienceDirect

Cleaner Environmental Systems

journal homepage: www.journals.elsevier.com/cleaner-environmental-systems

Exploring the use of a machine assisted goal and scope in a life cycle studies to understand stakeholder interest and priorities

Joseph Santhi Pechsiri [a,e,1,*], Alexandre Monteiro Souza [a,2,**], Chaveevan Pechsiri [b,3,***], Paulus Kapundja Shigwedha [c], Uasora Katjouanga [c], Benjamin Mapani [c], Rosa C. Goodman [d], Cecilia Sundberg [a], Niclas Ericsson [a]

[a] Department of Energy and Technology, Swedish University of Agricultural Sciences (SLU), SE 750 07, Uppsala, Sweden
[b] College of Innovative Technology and Engineering, Dhurakij Pundit University, Bangkok, 10210, Thailand
[c] Department of Civil, Mining, and Process Engineering. Namibia University of Science and Technology, 10005, Windhoek, Namibia
[d] Department of Forest Ecology and Management, Swedish University of Agricultural Sciences (SLU), Skogsmarksgränd, 901 83, Umeå, Sweden
[e] Department of Forest Biomaterials and Technology, Swedish University of Agricultural Sciences (SLU), Skogmarksgränd 17, 90183 Umeå, Sweden

ARTICLE INFO

Keywords:
Life cycle studies
Goal and scope
Stakeholder engagement
Natural language processing
Clustering
Machine learning
Impact categories and subcategories

ABSTRACT

The Goal and scope are essential phases within a life cycle study as they lay the foundation for the subsequent inventory modelling, impact assessment, and interpretation of results. Stakeholder engagement is critical throughout life cycle studies. Addressing diverse stakeholder interests and priorities have so far relied on stakeholder-expert dialogues, which remain challenging, particularly in projects with numerous stakeholders leading to a broad range of environmental, social, and economic impact categories and subcategories. This study therefore introduces a machine-assisted goal and scope approach to manage large volume of stakeholder responses generated in stakeholder-expert dialogues. It is designed to complement current manual stakeholder engagement approaches with semi-automated computer assisted analysis that identifies stakeholder interests, concerns, and prioritises them. We apply Natural Language Processing (NLP) in the goal and scope phase to preprocess stakeholder response documents collected during a life cycle study within a larger EU project. After preprocessing, unsupervised clustering algorithms were used to determine stakeholders' interests, concerns, and priorities. This innovative use of NLP and clustering was tested on a life cycle study of bioenergy value chains in Namibia (2021–2024). The approach successfully analysed stakeholder responses and identified key impact categories and subcategories on which to focus the assessment. Compared to manual methods, the machine-assisted goal and scope phase improved the level of detail while maintaining the same time frame and resource constraints. The current study serves as a proof of concept and demonstrates how life cycle studies can benefit from a machine-assisted goal and scope approach.

1. Introduction

Life cycle studies cover a range of methodologies to systematically quantify and assess the sustainability of a product or service. They take every step of the life cycle into account, including raw material extraction and upgrading, production, use, and end of life (waste treatment). There are separate methods to assess the three pillars of sustainability, such as environmental life cycle assessment (LCA), social LCA (SLCA), and life cycle costing (LCC). When all three are assessed together and quantitative and qualitative conclusions are drawn regarding potential trade-offs between and within the different pillars of sustainability, it is commonly referred to as a life cycle sustainability

* Corresponding author. Department of Energy and Technology, Swedish University of Agricultural Sciences (SLU), SE 750 07, Uppsala, Sweden.
** Corresponding author.
*** Corresponding author.
E-mail address: joseph.pechsiri@slu.se (J.S. Pechsiri).
[1] First author for machine-assisted scoping concept, environmental LCA, and LCC aspects of the paper.
[2] Second author for social LCA and stakeholder engagement aspects of the paper.
[3] Third author for machine learning and natural language processing aspects of the paper.

https://doi.org/10.1016/j.cesys.2025.100288
Received 11 November 2024; Received in revised form 19 March 2025; Accepted 30 May 2025
Available online 5 June 2025

assessment (LCSA) (Klöpffer, 2008). There are many ISO standards related to life cycle studies published under the ISO 14000 family (ISO). The main reference documents for LCA and SLCA are ISO 14040 (ISO, 2006a), ISO 14044 (Technical Committee ISO/TC 207, 2006) and ISO 14075 (ISO, 2024). LCC is less standardised and conceptually interpreted in different ways (Hunkeler et al., 2008). One way of including economic aspects in a life cycle study, coherent with the life cycle definition used in LCA and SLCA, is to perform an environmental LCC (Swarr et al., 2011).

Life cycle studies normally consist of 4 phases: a) goal and scope, b) inventory, c) impact assessment, and d) interpretation (Guinee et al., 2002). LCC is an exception. Since the inventory is described in monetary terms there is normally no need for a separate impact assessment step (Swarr et al., 2011). The goal of the study describes the reasons for carrying out the study, the intended application and audience, and whether the results will be disclosed to the public. In the scope, the system to be studied is described including its functions, the functional unit used, and the system boundaries. It also describes the allocation procedures, impact category selection, impact assessment methods, how these are interpreted, data requirements, assumptions, limitations, and whether and how a critical review will be performed (Technical Committee ISO/TC 207, 2006). The goal and scope phase can enhance the practical applicability of the study outcomes by directly involving project stakeholders (Guinee et al., 2002).

A rigorous life cycle sustainability assessment requires considering a broad set of impact categories. Selecting only a few categories risks overlooking burden shifting and undermines the goal of a comprehensive evaluation. While value choices and assumptions should be minimised, it is crucial to consider the interests, priorities, and preferences of the target audience and affected stakeholders, collectively referred to as interested parties in the SLCA standard (ISO, 2024). Their input can provide valuable insights, enabling meaningful conclusions and more relevant recommendations. Although methods that assess multiple impact categories are essential to prevent burden shifting, it is equally important to conduct a more detailed assessment of impact categories that are of particular concern to stakeholders. The selection of impacts for emphasis directly influences the level of detail required in the life cycle inventory and determines the depth of quantitative data collection necessary for a comprehensive assessment. Therefore, stakeholder participation and engagement are beneficial for ensuring the quality and usefulness of a life cycle study.

LCA and SLCA include different impact assessment methods, each with a broad and diverse set of impact categories The number of impact categories in a study depends on the chosen method and must align with the study's goal and intended application. However, there is limited guidance on prioritising impact categories in a way that ensures feasibility in later phases of a life cycle study (Rosenbaum, 2016), especially when multiple stakeholders are involved and the study targets a broad audience.

Time and resource constraints in any given life cycle study makes it challenging to collect detailed data and assess all possible impact categories, regardless of the availability of background data in life cycle databases. This challenge is particularly pronounced in studies covering the social dimension, where stakeholder preferences play a key role in guiding the assessment (ISO, 2024), (UNEP, 2020), (UNEP, 2021). To enhance the relevance and usefulness of the study, stakeholder interest and priorities should inform the selection of categories and subcategories on which to focus data collection and impact assessment efforts, ensuring more meaningful and useful study outcomes.

Depending on the purpose and scope of a life cycle study, the number of stakeholders and the size of the target audience can vary significantly. In small studies with a specific target group, selecting relevant impact categories and interpreting the study is often straightforward (Curran, 2016). However, when assessing larger systems and value chains with numerous stakeholders - such as policy makers, industries, non-state actors, local communities, workers, and unions - stakeholder interests and concerns become more complex and sometimes contradictory. This complicates the goal and scope phase in general, particularly the selection of assessment methods and impact categories.

Moreover, when practitioners engage with stakeholders, maintaining complete objectivity is challenging, even though it is essential for conducting an unbiased analysis. This not only affects the outcome of the study, but also shapes how the system is perceived by those who receive the practitioners interpretation (Yamamoto, 2012). Additionally, stakeholder responses may be unintentionally influenced by the mere presence of practitioners, potentially leading to a systematic misrepresentation of their preferences and priorities (Miyazaki and Taylor, 2008). Given these challenges, there is a need for structured methods that improve transparency and minimise subjectivity in stakeholder engagement data analysis. This study explores the application of natural language processing (NLP) and clustering techniques to address these issues.

Life cycle practitioners often simplify the interpretation of large datasets generated from multiple impact categories and diverse stakeholder groups by using scoring and indexing approaches, (e.g. (Romagnoli et al., 2024), (Abdella et al., 2020), (Bachmann, 2013)). While these methods aid in result interpretation and communication, they may lack the depth needed to address concerns that are highly relevant to specific stakeholder groups, particularly smaller or more localized ones. As a result, the study's findings may not be relevant to certain stakeholders (Guinee et al., 2002). To enhance the relevance of a life cycle study's results, this study explores the use of NLP tools (Chowdhury, 2003), and machine learning (ML) or data mining techniques (e.g. a clustering algorithm as an unsupervised method in ML (Yadav and Sharma, 2013)) to identify stakeholder interests and priorities during the goal and scope phase of a life cycle study. This is especially relevant to life cycle studies that include an SLCA component, where stakeholder perspectives play a central role.

This study aims to demonstrate how the NLP and clustering techniques can be combined and applied to analyse stakeholder responses to aid in the identification of impact assessment categories and subcategories of high relevance to stakeholders during the goal and scope phase of a life cycle study. This application has not been thoroughly explored in the life cycle methodology related literature.

The proposed approach was applied and tested in a larger EU project, where a complete life cycle sustainability assessment (LCSA) was carried out to assess the potential environmental, social, and economic impacts of introducing potential new biomass-based value chains in southern Africa. The study discusses how this machine-assisted approach can enhance the goal and scope phase and improve the relevance of the study's outcomes for stakeholders and the target audience. By automating aspects of stakeholder response analysis, this study addresses the challenge of maintaining objectivity by the life cycle practitioner when identifying impact categories of high relevance and concern to stakeholders.

2. Materials and methods

2.1. The proposed machine assisted goal and scoping concept

ML is a computational approach that enables computer models to autonomously learn from data and perform complex tasks through the use of machine learning tools without explicit programming. This learning process can involve acquiring new declarative knowledge, developing motor and cognitive skills through instruction or practice, discovering new facts and theories through observation and experimentation, or organising existing knowledge into structured insights (Carbonell et al., 1983). ML is often described as an artificial mechanised system that refines its performance through experience (Mitchell, 2006) and can be integrated with data mining techniques to enhance analytical effectiveness (Dogan and Birant, 2021).

NLP, a subfield of artificial intelligence, focuses on the

computational understanding and manipulation of the human language in text or speech form (Chowdhury, 2003). In this study, NLP is applied to process text-based responses from stakeholders, which are collected during the goal and scope phase of the life cycle study. A clustering technique is then employed to group stakeholder interests and priorities, helping to identify impact categories that require emphasis. These identified impact categories subsequently inform the life cycle inventory, impact assessment and interpretation phases, following an iterative life cycle study methodology (Technical Committee ISO/TC 207, 2006).

While the application of ML and NLP techniques in life cycle studies are still emerging, their adoption has grown in recent years (Romeiko et al., 2024). To date, ML has been predominantly used in life cycle inventory modelling, impact assessment, and interpretation phases (Romeiko et al., 2024), (Ghoroghi et al., 2022). These applications primarily aim to address data gaps (Meng et al., 2019), manage uncertainties (Abokersh et al., 2020), and improve life cycle studies through systems optimisation (Azari et al., 2016), hotspot identification (Zhao et al., 2019), and sensitivity analysis (Abokersh et al., 2020). NLP has also been applied to incorporate textual data into life cycle studies, further supporting data integration and analysis (Chiu et al., 2024). Despite the widespread use of ML and NLP in industrial applications (Pechsiri and Kawtrakul, 2007), (Pechsiri et al., 2016), their integration into the goal and scope phase of life cycle studies remains largely unexplored.

Including stakeholders in the goal and scope phase of a life cycle study is important for ensuring that the assessment aligns with their priorities and concerns. Stakeholder perspectives provide crucial input for drawing meaningful conclusions and making recommendations during the life cycle interpretation phase. Traditionally, participatory approaches involve stakeholder engagement tools, where life cycle practitioners manually process the collected input to define the study's goal and scope (Fig. 1, approach A). However, this manual approach can be time-intensive and subject to expert interpretation biases.

This study introduces an alternative machine-assisted approach. ML and NLP were incorporated to enhance the goal and scope phase (Fig. 1, approach B). By leveraging computational tools, the proposed method aims to automate the alignment of the study's goal and scope with stakeholder priorities, improving efficiency and reducing subjectivity.

In our proposed machine assisted goal and scope approach, stakeholder concerns and priorities were collected as unfiltered and unscreened textual data. This approach encourages the integration of diverse stakeholder perspectives, which can improve the relevance of the final study (Sala et al., 2013). To prevent any unintended influence on stakeholder perspectives (Miyazaki and Taylor, 2008), only open-ended questions were used rather than multiple-choice or structured survey questions.

Stakeholders were asked broad, qualitative questions such as.

- "What are your social, economic, and/or environmental concerns at the moment?"
- "What are your interests and concerns within the biomass industry?"

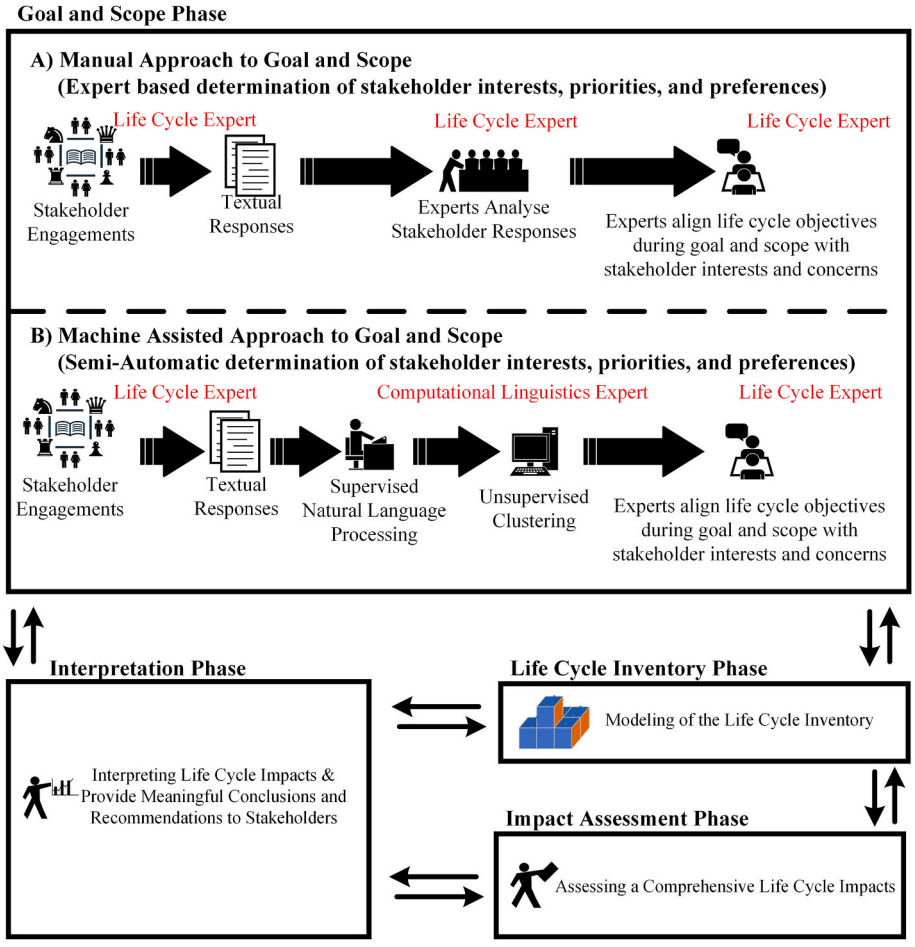

Fig. 1. The four methodological phases of a Life Cycle Study. Traditionally, life cycle practitioners manually analyse stakeholder inputs during the goal and scope phase (A) using various participatory tools. This study proposes a modified approach (B), which integrates machine assistance through natural language processing and machine learning. In the second approach, computational linguistics experts process stakeholder responses to help semi-automatically identify stakeholder interests, concerns, and preferences, which are then aligned with life cycle study objectives by life cycle practitioners.

These responses were collected into a textual data or corpus for subsequent processing using NLP and ML techniques.

When applying ML and NLP in decision-support systems, it is common to include a supervised NLP phase to process textual input (Pechsiri et al., 2016), (Pechsiri and Piriyakul, 2021), (Pechsiri and Piriyakul, 2016a). In this study, NLP techniques were used to transform stakeholder responses into structural linguistic data, ensuring that key concerns and priorities could be systematically analysed. The NLP process included lemmatisation, which refines textual data by reducing words to their base forms while preserving their intended meaning. Information identifying the sources was removed to minimise potential bias from life cycle practitioners in subsequent stages and to comply with General Data Protection Regulation (GDPR) requirements (The European Parliament and of the Council).

The processed textual data was then used to extract key-term features, which served as input for the next stage: clustering. In this step, linguistic tags were removed, and stakeholder responses were grouped into three arbitrary clusters based on shared features. These clusters represented broad categories of stakeholder interests and concerns, which served as a foundation for prioritising impact categories in the life cycle study.

Two main common clustering techniques were applied (Jung et al., 2014), (Schön, 2009), (Krantz et al., 2009) to group stakeholder concerns.

1. k-means clustering algorithm, which is based on Euclidean distance measurements.
2. Expectation Maximization (EM) clustering algorithm, which employs probabilistic statistical modelling to iteratively optimise parameter estimation.

K-means clustering is widely used to identify homogeneous clusters within datasets based on predefined characteristics of interests (Krantz et al., 2009), (Schreiber and Pekarik, 2014). In previous life cycle studies, k-means has been combined with logistic regression and weighted scaling of social, economic, and environmental values to simplify complex sustainability assessments. For example, k-means were used to aggregate life cycle impact assessment results into a composite sustainability index for stakeholders and complementing multi-criteria decision-making methods (Abdella et al., 2020). In this study, we used k-means clustering differently. Inspired by its application in audience segmentation studies (e.g understanding museum visitors' perspectives (Krantz et al., 2009)), we applied K-means clustering to identify patterns in stakeholder concerns and priorities from the supervised NLP processed textual data. Unlike studies that relied on predefined stakeholder interests, this study predefined three arbitrary clusters as a demonstration of how clustering can assist in identifying key stakeholder interests. This choice of three clusters served as an initial proof of concept, allowing for an exploration of how unsupervised clustering can structure stakeholder concerns in the goal and scope phase.

As an alternative approach, EM clustering was also applied in a similar manner. Inspired by Darena et al. (2012), EM clustering was explored to compare its effectiveness against k-means, providing an additional perspective on how different clustering techniques can be leveraged to enhance stakeholder analysis in life cycle studies.

After the clustering step, the general concept of each cluster was interpreted or assigned a theme label by life cycle practitioners to represent each cluster as a group class. Unlike typical classification models, which rely on predefined labels and performance metrics (such as confusion matrices, precision, recall and false positives), this study focused on expert-driven thematic assignment to help ensure that the identified stakeholder concerns aligned meaningfully with the life cycle study objectives. As a result, traditional performance evaluation metrics were not applicable.

A panel of three life cycle practitioners examined the key-term features within each group to assign the theme label that would best represent the primary stakeholder concerns captured in the responses. To minimise subjectivity bias, this thematic assignment of clusters was conducted without knowledge of who the stakeholders were. By redacting stakeholder sources and only presenting the lemmatised words from each cluster, this method enables experts to make interpretations based solely on the semantic patterns of stakeholder responses, rather than any preconceived notions about specific stakeholder groups.

Once themes were assigned, life cycle practitioners identified impact categories of high relevance to be used as guidance for subsequent phases of a life cycle study. The assignment of impact categories was determined by analysing which key-term features appeared most frequently in each cluster, reflecting stakeholder priorities and concerns. These impact categories then informed the life cycle inventory, impact assessment, and interpretation phases to help the assessment remained stakeholder focused and contextually relevant. The structured assignment of themes and impact categories also helps translate stakeholder concerns into meaningful recommendations for decision-making in the interpretation phase (Fig. 1).

The machine assisted goal and scope approach described here was tested alongside a manual approach as part of the larger project's life cycle studies, allowing for an assessment of its limitations and benefits.

2.2. Case study

This case study was conducted in Namibia between the years of 2021 and 2024 as part of a broader effort to assess the sustainability of biomass-based value chains. Bush encroachment is a significant environmental and economic challenge in the region (O'Connor et al., 2014), and the project aimed to investigate the feasibility of establishing value chains that would utilize unwanted woody biomass while providing economic incentives for sustainable land management. Extensive stakeholder engagements were conducted to ensure that the life cycle study reflected local environmental, social and economic concerns.

Life cycle studies were used to assess the environmental, social and economic sustainability of the biomass value chains. The assessment made use of the product environmental footprint (EF3.1) and the Ecoinvent 3.7.1 database (Andreasi Bassi et al., 2023), (Crenna et al., 2019), (Wernet et al., 2016), the product social impact life cycle assessment (PSILCA) database (UNEP, 2020), (Kirill et al., 2020), and the net added value approach based on primary and literature data (Swarr et al., 2011), (Damodaran, 2017) for the environmental, social, and economic assessments, respectively.

2.2.1. Stakeholder engagements and participatory methods

The stakeholder engagement process followed four key phases: (1) identifying relevant value chain stakeholders with local experts, (2) selecting and preparing participatory activities, (3) conducting stakeholder engagement sessions, and (4) processing stakeholder input for integration into the life cycle study (Fig. 2). This structured approach ensured that stakeholder concerns were systematically incorporated into the assessment.

2.2.1.1. Stakeholder groups. Stakeholder mapping was conducted through literature review (Mlunga and Gschwender, 2015), (Brüntrup et al., 2012) and consultations with project partners to identify relevant groups. The stakeholder groups engaged in the study included government agencies, value chain operators, unions, farmers, industrial off-takers, households and local communities (Table 1). Representatives of these groups were engaged throughout the project and, additional stakeholders were identified through an evolving engagement process. However, demographic data such as gender distribution, age range, and educational background were not systematically recorded. This study was designed as a proof of concept, focusing on testing the

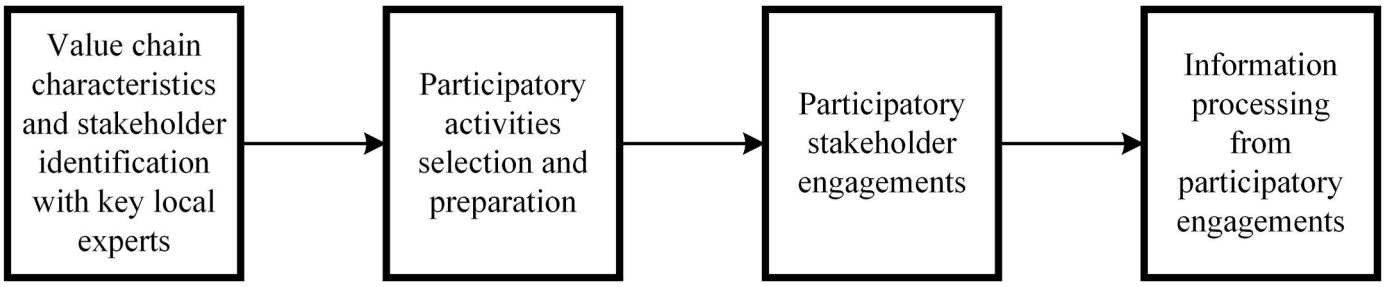

Fig. 2. Work flow of stakeholder participation planning and execution.

Table 1
Groups and types of stakeholders that were engaged in Namibia in the case study.

Stakeholder group	Type of stakeholder
Governmental agencies	Government Ministries
	Academic Institutions
	Healthcare Services in Namibia
Non-Governmental Organizations (NGOs)	Developmental Organisation
	Conservation Organisation
Unions and Associations	Unions representing the agriculture sector
	Unions representing workers and industries
Harvest Operators	Manual Harvesting Operators
	Mechanised Harvesting Operators
Potential large and small-scale commercial end users	Industries
	Small-Medium Enterprises
	Households
Commercial farmers	Cattle Ranges
	Charcoal producers
Tourism Service Providers	Game farms
	Accommodation providers
Smallholders from communal areas	Communal areas of Namibia
Financial Institution	Consumer Banking Service Provider
Value Chain Operators	Biomass Industry Service Providers
	Certification Service providers

machine-assisted approach rather than analysing demographic influences on stakeholder concerns. While demographic insights could provide additional depth in future research, their absence does not affect the primary objective of evaluating the feasibility of the proposed method.

Partner organizations facilitated access to many stakeholders, and participation levels varied across groups. Communal area stakeholders had higher participation due to their geographic proximity of individuals to their community centres. This made group activities easier to organise, especially when community leaders were involved, but low literacy and language barriers posed challenges. Commercial farmers, were geographically dispersed, which limited group participation. Some farmers also had engagement fatigue from previous development projects, leading to lower attendance. Unions, associations, universities and governmental agencies participated through their representatives.

2.2.1.2. Selection of participatory tools. We utilised different participatory tools depending on the group and structure of stakeholders in each engagement event. The selection of participatory tools was based on.

1. Stakeholder group size - large groups required interactive methods such as brainstorming.
2. Target information - Specific concerns required structured discussions.
3. Available resources. - Methods were adapted to time and logistical constraints.

The study employed brainstorming, semi-structured dialogues, focus groups, and formal surveys to collect stakeholder input. These methods were chosen to maximise inclusion and relevance, ensuring broad representation of stakeholder concerns.

2.2.2. Implementation of stakeholder engagement methods

Brainstorming sessions was used primarily with groups in communal areas to gather concerns and priorities. participants were given note cards to write their primary concerns, particularly regarding bush biomass harvesting. Each note card contained a single concern, while each individual was allowed to write as many note cards as needed, ensuring that a diverse set of issues were captured. This method allowed stakeholders to express their views independently, minimising the risk of dominant voices overshadowing others (Geilfus, 2008).

Semi-structured dialogues were conducted with individuals, families, or small groups. This method facilitated in-depth conversations and allowed researchers to explore concerns beyond predefined topics. Unlike formal questionnaires, semi-structured dialogues allowed stakeholders to introduce new discussion points, which reduces bias from preset questions (Geilfus, 2008).

Focus groups, a variation of semi-structured dialogues, were conducted with more groups focused on specific activities. For example, a workshop with farmers involved in bush thinning was held to understand their perspectives on biomass harvesting.

Formal questionnaires were used to engage a broader set of stakeholders, particularly those unable to attend in-person discussions. The questionnaires included open-ended questions to capture stakeholder perspectives on social, economic, and environmental values related to the biomass industry. Due to time and resources constraints, formal questionnaires were selectively distributed. One key data collection event was the 2023 Biomass Fair in Namibia, where stakeholders from the biomass sector participated in surveys. Fig. 3 provides an overview of the stakeholder groups and their associated biomass-related activities.

2.3. Manual approach for identifying relevant impact categories

To systematically categorise stakeholder concerns, the stakeholder input collected through participatory methods was analysed using two complementary approaches. The manual approach relied on expert interpretation to sort stakeholder responses into thematic categories linked to relevant impact categories. This method provided a baseline for assessing the feasibility and limitations of machine assisted processing.

2.3.1. Data collection, categorisation and impact category assignment

Stakeholder responses were collected through the participatory tools described in section 2.2.1 and compiled into a corpus for the manual and machine assisted analyses (Supplementary material A). The manual approach required life cycle practitioners to identify patterns in the responses and organise concerns in thematic clusters. These clusters were then linked to impact categories commonly used in different types of LCA, SLCA and LCC.

A matrix was developed to structure stakeholder concerns by environmental, social and economic dimensions. This matrix supported the

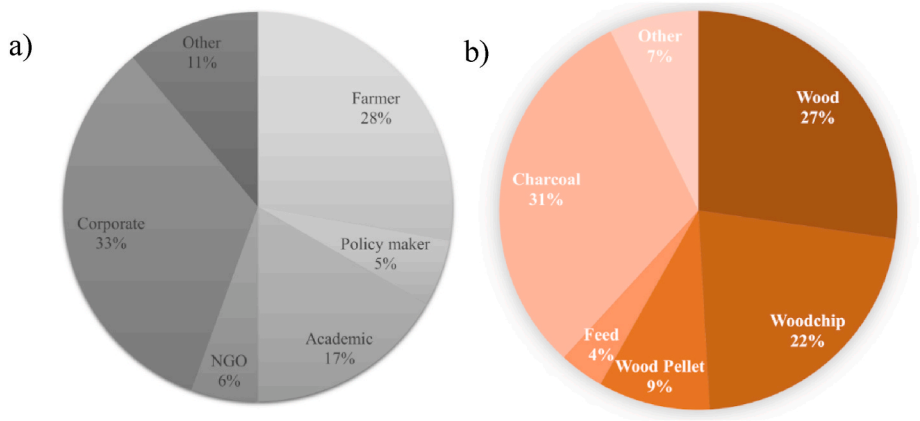

Fig. 3. Grouping of stakeholders that responded to the questionnaire at the 2023 Biomass Fair (a) and the biomass products engaged with by these stakeholders (b).

systematic prioritisation of impact categories, ensuring that key stakeholder concerns were reflected in the life cycle study. It also served as a reference for aligning stakeholder priorities with system modelling and data collection in the inventory phase.

The assignment of impact categories was guided by: (i) the frequency and intensity of concerns raised by stakeholders, (ii) relevance to

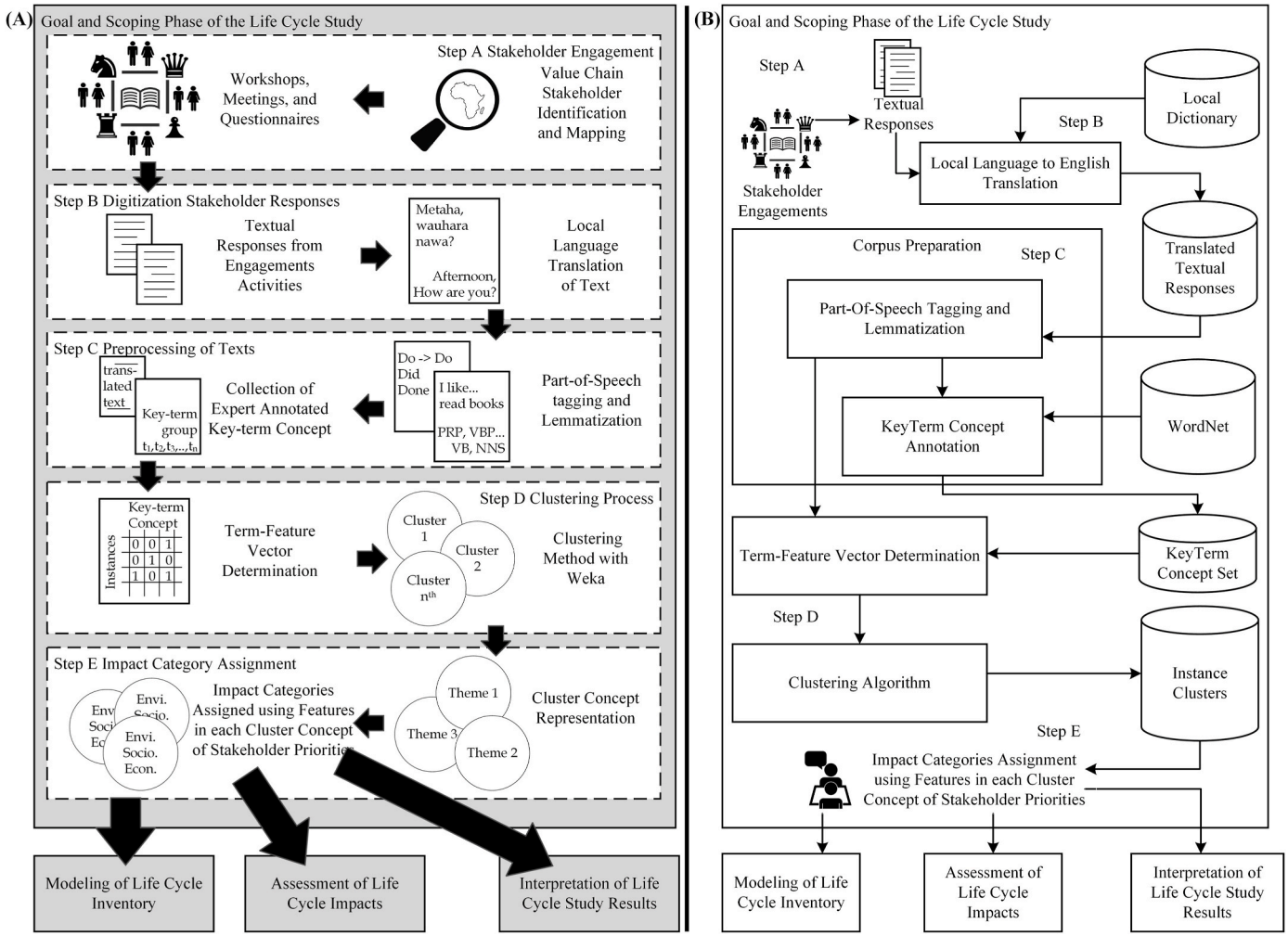

Fig. 4. Two systematic representations of processes: a) illustrates the sequential workflow used to process stakeholder concerns, from data collection to impact category assignment, while b) presents the same process from a computational systems perspective, showing the data storage (as cylinders), processing flow (as boxes), and algorithmic structure necessary for automation Stakeholder responses were collected in engagement activities (step A), translated (step B), and processed using natural language processing techniques (step C). Clustering was performed using the Waikato Environment for Knowledge Analysis platform (WEKA) which found arbitrary clusters of stakeholder interests and concerns (step D). These clusters were then assigned themes by LCA experts following their respective features within each cluster and used to identify high priority impact categories and subcategories (Step E). Steps A and E were conducted manually by LCA experts (authors); Steps B and C were conducted using machine assistance under LCA expert supervision (authors); Step D was unsupervised.

existing LCA, SLCA and LCC impact assessment methods, (iii) cross-cutting themes affecting multiple stakeholder groups (e.g. economic viability, environmental degradation, or social well-being. Responses were recorded with accompanying metadata to maintain traceability and provide context for impact category assignment. Since some responses represented individuals and, others reflected collective perspectives (e.g. community representatives, NGOs) the data was not processed statistically but rather qualitatively categorised based on thematic clustering.

While expert interpretation attempts to capture contextual accuracy, the process is time-intensive and subjective. Potential limitations of the manual approach included (i) inconsistencies in expert judgment across different life cycle practitioners, (ii) difficulty in processing large volumes of stakeholder input efficiently, and (iii) potential bias in assigning impact categories based on expert perceptions rather than purely data-driven insights. The machine-assisted approach was therefore developed to increase transparency, efficiency and replicability in identifying impact categories of high relevance and priority to stakeholders.

2.4. Machine-assisted approach for identifying relevant impact categories

The machine-assisted approach used the same transcribed stakeholder responses as the manual approach (Supplementary material A). To facilitate structured processing, semi-automatic coupling of NLP and clustering techniques were applied.

The workflow consists of five key steps (Fig. 4a). Step A and B consist of data collection and transcription. In the case study this is where the stakeholder input was gathered using participatory methods and compiled into a text corpus. Since responses were provided in various local languages and dialects, low literacy and language barriers posed significant challenges and local translators were often required.

A total of 114 responses with 1477 words were collected during the stakeholder engagements. Given that many local dialects lack written standardization, a professional translator assisted in ensuring accuracy and consistency before text processing.

In Step C (Fig. 4), the collected data are digitalised, stored, and preprocessed using NLP with techniques such as part-of-speech tagging, lemmatisation and key-term annotation. Part of speech tagging is a linguistic method that refers to markup tagging of words based on their grammatical part of speech, e.g. “I eat carrots” is tagged as “I/PRP eat/VBP carrots/NNS”, where PRP, VBP, and NNS refer to personal pronoun, verb present, and plural noun, respectively. During this step, words are assigned syntactic categories (Fig. 5) using Penn Part of Speech tags (Santorini, 1990). Lemmatisation reduces words to their base form. As an example, words like “go”, “went”, “gone” are all related to the word “go”. In the case study, we used a lemmatiser from the Centre for Language Technology at University of Copenhagen (Jongejan and Dalianis, 2009), (Jongejan and Haltrup, 2005).

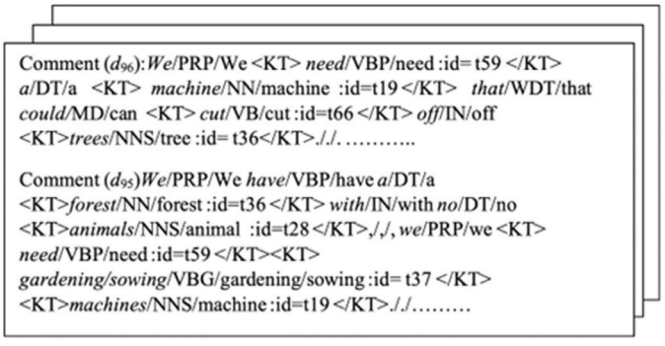

Fig. 5. An example of POS tagged (e.g. PRP, NN, VBG) and annotated KT tags used for annotating key terms as key-term concepts in textual documents containing stakeholder responses collected during stakeholder engagement activities.

Key-term annotation identifies domain-specific terms based on a lexical database, WordNet ((Fellbaum, 2010), (Miller, 1995)), in order to facilitate the upcoming clustering process. In the case study, each identified key-term concept (t_i; i = 1,2, …,*numberOfKeyTermConcerpts*) was added to a key-term set (T) (equation (1)):

$$T = \{t_1, t_2, t_3, \ldots, t_{\text{numberOfKeyTermConcepts}}\} \quad (1)$$

The key-term set was then used to extract key-term concept features from the corpus, to be used as clustering features. The annotations of key-term concepts was achieved using computational linguistic experts' annotation of key-term concepts based on WordNet (Fellbaum, 2010), (Miller, 1995) with a random sample of the corpus. An example of 50 random textual responses, with annotated key-term concepts can be found in Supplementary Material B.

Step D is the clustering process. The term-feature vectors within the Term-Feature Vector Determination step (Fig. 4a) are initially determined from the corpus. The feature vectors are extracted before using clustering techniques to cluster responses into arbitrary groups, representing stakeholder concerns and priorities. A term-feature vector ($\mathbf{tv}_{dj}$) is a binary feature vector of the extracted key-term concept from textual responses or document instances (d_j; j = 1,2, …,*numberOfDocuments*) on the corpus. Each binary feature vector is a numerical representation of textual data that enable the posterior clustering algorithms to process and analyse the data. In the case study, $\mathbf{tv}_{dj}$ is determined from each stakeholder textual response by extracting the presence of key-term concepts in d_j. String matching method ((Santorini, 1990), (Pechsiri et al., 2020)) was used between each T element and each d_j terms (Table 2 and Supplementary Material C).

Once the feature vector had been determined, the k-means and EM clustering techniques were applied. The two techniques algorithms were performed in parallel to demonstrate the potential use for different clustering techniques in step D.

The k-means clustering was performed using data mining software from the Waikato Environment for Knowledge Analysis platform (WEKA) (Bouckaert et al., 2016), (Hall et al., 2009). In the k-means clustering algorithm, n samples (number of instances of stakeholder textual responses) were analysed using k-means instance clustering (Equation (2)) (Pechsiri and Piriyakul, 2016b), (Aloise et al., 2009).

$$\text{Cluster}(\mathbf{tv}_{dj}) = \arg\min_{1\le k\le K} \left\|\mathbf{tv}_{dj} - \mu_k\right\|^2 \quad (2)$$

where $\mathbf{tv}_{dj}$ is the term-feature vector of t_i extracted from d_j; i = 1,2,.., *numberOfKeyTermConcept*; j = 1,2,..,*numberOfDocuments*; *numberOfDocuments* is the number of documents in the corpus; μ_k is the mean vector of the cluster.

In Eq. (2), k represents the number of clusters used as representatives for the different cluster concepts from stakeholder textual responses, i is the number of key-term concepts, j is the number of documents in the main corpus, and μ_k is the mean vector of the cluster k. The number of clusters is predefined between 2 and 10 (see Supplementary Material D). The centroid coordinate values resulting from each clustering of key-term features was used to determine a final k-value. A k-value of 3 was chosen to generate three clusters of stakeholder responses. This choice was made pragmatically based on expert judgment to illustrate the feasibility of applying clustering techniques in a life cycle study. Supplementary Material D provides additional examples with alternative k-values.

Table 2
Example of vectors determined from stakeholder textual responses.

Instances or Documents	t_1	t_2	…	t_{65}	…	$t_{numberOfKeyTermConcepts}$
d_1	0	0	…	0	…	…
d_2	0	0	…	1	…	…
….	…	…	…	…	…	…
d_j	0	1	…	0	…	…

Another common clustering algorithm applied in the case study is EM clustering (Schön, 2009), (Dempster et al., 1977). This was conducted using WEKA to observe how the key-term feature concepts affect each cluster. Unlike Euclidean distance based k-means, EM clustering uses statistical methods to determine clusters (Jung et al., 2014). There are two sub-processes in the EM clustering algorithm (Schön, 2009), (Dempster et al., 1977). The first sub-process guesses initial parameters: mean and standard deviation (if using normal distribution model). The second sub-process iteratively refines the parameters with Expectation and Maximization steps. In the Expectation step, the membership possibility for each instance is computed based on the initial parameter values. In the Maximization step, the parameters are recomputed based on the new membership possibilities.

Once the three arbitrary clusters had been generated from the key-term features in the case study, the response instances containing the key-term features in each cluster were documented as instance clusters. These were the final outputs from the NLP and clustering processes.

Step E is the impact category assignment and consists of interpreting the clustered stakeholder responses, assigning thematic labels and matching them to relevant life cycle impact assessment categories. Once the clusters containing the key-term features were obtained through unsupervised clustering a panel of 3 life cycle practitioners reviewed and categorised the stakeholder concerns and priorities Each expert had a different professional and national background and came with different experiences in life cycle studies, thus giving a diverse range of perspectives.

To assign themes, the experts analysed key-term features within each cluster and reached a consensus through a majority voting process (max-win voting). Thematic labels were then used to identify impact categories of high relevance to the stakeholders. If consensus could not be reached for a given cluster, the clustering results were considered inconclusive and a different clustering method was attempted.

For quality control, k-means and EM clustering were applied in parallel, as a cross-validation measure. The final themed clusters (representing the three primary stakeholder concerns) were then mapped to impact categories in LCA, SLCA and LCC. Ultimately, the thematic clusters reflected stakeholder priorities, guiding the selection of impact categories which warranted particular focus when planning data collection and assessment. This structured semi-automated approach helped mitigate bias and promote a stakeholder focused life cycle study while maintaining methodological rigour.

3. Results

Both the manual and machine assisted goal and scope approaches successfully identified stakeholder concerns and associated impact categories, but the manual approach introduced the risk of subjectivity through bias and inconsistency in interpreting the responses by life cycle practitioners. The machine assisted approach aimed to mitigate these limitations by using automating pattern recognition and categorisation, and succeeded in reducing the dependency on expert judgment.

3.1. Manual approach

The result of manually identifying prioritised impact categories is shown in Table 3. They are illustrated using a matrix of impact categories and subcategories identified as highly prioritised by different stakeholder groups, based on their responses during participatory activities.

The SLCA impact subcategory ‘local employment’ was the only one prioritised by all stakeholder groups (Table 3). This may be due to the fact that Namibia has been afflicted by high unemployment in recent years (Amutenya, 2021). Most stakeholder groups had the category ‘access of stakeholders to resources and basic services’ as a serious concern. This was often connected to the need for either infrastructure or educational services, especially in rural areas (Nguvenjengua and Undji, 2017). ‘Climate impact’ was highlighted by several stakeholder groups. As a response to extreme poverty and income inequality (Nguvenjengua and Undji, 2017), (NPC, 2023), stakeholders directly and indirectly involved in the value chains (e.g. value chain operators and harvest operators, and government agencies and academics) prioritised ‘cost’ and ‘added value’ in the value chain as a whole from an LCC perspective. Some stakeholders raised issues related to ‘biodiversity’, particularly concerning the species of animals and vegetation that could be affected by bush related activities. Some stakeholder groups prioritised ‘fair salary’, ‘human toxicity’, ‘health and safety’, ‘work equality’, and ‘water footprint’ for points of discussion, but these were not considered as important as employment.

It is worth noting that Table 3 was solely based on the issues raised during participatory activities, and does not represent all the potential impacts of the investigated value chains. Many real impacts are often overlooked by stakeholders. For example, healthcare service providers (included in the ‘government agencies’ stakeholder group) identified ‘risk to injuries’ and ‘development of respiratory illnesses’ as potential impacts on workers in the bioenergy value chains. These were, however, frequently downplayed by these stakeholder groups themselves. The lack of concern for these issues could partly be explained by under-equipped and underfunded healthcare service providers, leading to a lack of awareness of the issues amongst other stakeholders. This highlights the need and relevance to include a comprehensive set of impact

Table 3
Manually determined impact categories, based on stakeholder responses (Supplementary Material A).

Impact Categories and Subcategories Raised	Engaged Stakeholders						
	Government Agencies	NGOs	Small Holders in Communal Areas	Farmers & Harvest Operators	Value Chain Operators	End users	Academics
Energy Demand			■		■		■
Resource Depletion			■	■		■	
Climate impacts			■	■		■	■
Particulate Matter			■	■			
Water footprint			■				
Human Toxicity				■			
Biodiversity			■	■			■
Economic Costs				■	■	■	■
Added value	■		■	■	■	■	■
Resources and Services Access		■	■	■	■	■	
Employment	■	■	■	■	■	■	■
Working Hours							■
Fair salary				■			■
Health and Safety				■			
Women at work and Equality					■		

categories in a life cycle study, which is recommended by guidelines and standards (ISO, 2024), (Rosenbaum, 2016), (ISO, 2006b). Nevertheless, the manual approach identifies impact categories that are of heightened importance by stakeholders to be used further in the data collection, impact assessment and interpretation phase of a life cycle study.

A key challenge with the manual approach, especially under time and resource constraints, is the risk of subjectivity in how practitioners interpret responses, first as issues and subsequently into impact categories. In one of the engagements performed in this study, one of the respondents mentioned just one word: 'Water'. This can be interpreted in at least three different ways: a) access to drinkable water and sanitation; b) water resource scarcity in regions surrounding the Kalahari Desert; and c) the effect of water on the balance between woody vegetation and grass in the region. When translating this into impact categories, it could be interpreted as a need to emphasize the 'water footprint', or 'biodiversity impacts' in an LCA; the need to emphasize 'safe and healthy living conditions for local communities' in the SLCA; focusing on the subcategories relating to 'access to resources and services' (e.g. drinking water coverage and sanitation coverage); or all of the above. The issue could be resolved through an iterative process with more stakeholder interaction. However, given limited resources and with time constraints, there is a risk that additional rounds of stakeholder engagement activities may not take place, and the selection of high priority impact categories may become a subjective choice.

3.2. Machine-assisted approach

In the machine-assisted approach, NLP techniques were combined with clusters to determine stakeholder concerns and priorities during the goal and scope phase of the life cycle study. Using WordNet ((Fellbaum, 2010), (Miller, 1995)), 75 annotated key-term concepts were extracted from the stakeholder responses (Table 4, Supplementary Material B).

Once extracted, feature vectors were generated for the key-term concepts (see supplementary material A), enabling further analysis. Two different clustering methods, k-means and EM were applied in parallel to test their effectiveness in identifying distinct patterns within stakeholder concerns and priorities.

3.2.1. K-means clustering and expert-assigned themes and impact categories

The k-means clustering method, using k = 3, was applied following Eq. (2) using the WEKA software. The detailed results of the clustering analysis, including the centroid coordinate values, are provided in Supplementary Material D. A subset of these values is presented in Table 5, which illustrates the three clusters along with their associated key-term concepts.

Term-feature vectors were categorised based on their degree of association with each cluster. Terms with a feature vector score below 0.1 were considered marginally associated, while those with a score exceeding 0.2 were considered highly associated with a given cluster. Scores falling between 0.1 and 0.2 were considered moderately associated. (Supplementary Material E contains the resulting clusters).

Cluster 1 contained the largest proportion of feature vectors, representing 80.5 % of the total feature vectors, while Cluster 0 accounted for 14.5 % and Cluster 2 contained 5 % of the feature vectors (Table 5). Certain key-term concepts (e.g t8, t65, t66) were highly associated with Cluster 2, whereas others (t7, t8, t65) were moderately associated with Cluster 1 (For a full list of centroid coordinate values, refer to Supplementary Material D). The results indicated that most stakeholder responses could be distinctly classified into a single cluster, facilitating the subsequent thematic analysis by life cycle practitioners.

Table 4
Examples of key-term concept features (see Supplementary Material A).

Key Term ID	Annotated Key-Term Concept
t1	GovermentOfNamibia2/regionalOffice/Namibia/town/Okakarara1/Grootfontein1
…	…
t4	environmentManagement/GreenHouseGas/ClimateChange/rainfall/healthyAir
…	…
t75	Living/betterLiving/livelihoods

Table 5
Example of centroid coordinate values for key-term concepts derived from stakeholder textual responses within each cluster, based on the k-means clustering method ($k = 3$) in WEKA (Supplementary Material D).

Key-Term ID (from Table 4)	If All Key-Term were considered as a group (100 %)	Key-Term in Cluster 0 (14.5 %)	Key-Term in Cluster 1 (80.5 %)	Key-Term in Cluster 2 (5.0 %)
…	…	…	…	…
t7	0.1068	0.0667	0.1205	0
t8	0.165	0.1333	0.1687	0.2
…	…	…	…	…
t65	0.1845	0.2667	0.1566	0.4
t66	0.0485	0.0667	0.0361	0.2
…	…	…	…	…

To interpret the clustering results, life cycle practitioners manually assigned thematic labels to each cluster (Supplementary Material E, F, and G). The identified themes were then aligned with existing impact categories to identify impact categories of high relevance and priority to stakeholders (Table 6).

The results indicated that stakeholder responses could be grouped into three primary themes.

1. local community and smallholder concerns,
2. issues of employment and decent work, and
3. future technology and education requirements.

As responses did not overlap between clusters, the assignment of thematic labels by the life cycle practitioner panel was conducted with high consistency and efficiency. Based on the themes, the impact categories and subcategories of highest relevance and priority to the stakeholders included: land transformation, climate impact, water footprint, energy demand, toxicity, access to services by small holders, access to material resources by local communities, employment qualities for workers, fair competition in the value chain, capital expenditures, and potential financial performance.

3.2.2. EM clusters and expert-assigned themes and impact categories

To compare the k-means clustering results with an alternative technique, EM clustering was applied using the same data. Unlike k-means, which assigns each response to a single cluster, EM clustering assigns responses probabilistically, allowing overlapping membership of responses across multiple clusters.

The application of EM clustering resulted in a more even distribution of responses across clusters than k-means clustering. Several of the responses were assigned to multiple clusters using EM clustering, where Cluster 1 contained 40 % of the feature vectors, while Clusters 0 and 2 contained 35 % and 25 % of the total feature vectors respectively.

Since some stakeholder responses were present in multiple clusters, the resulting stakeholder concerns covered by the themes assigned to each cluster overlapped significantly. This affected the expert interpretation process. It took about twice as long to reach consensus on thematic labels using EM clustering, compared to k-means clustering.

The three main themes identified using EM clustering were (Table 7).

1. economic developmental challenges,
2. land management funding and investments, and

Table 6
High priority impact categories and subcategories for each k-mean based cluster of responses and features.

Cluster	Cluster theme	LCC Impact Categories	Environmental LCA Impact Categories	Social LCA Impact Categories and Subcategories
Cluster 0	Local Communities and Small holder concerns	- Revenue - Cost	- Land Transformation - Water Footprint - Resource Depletion	**Small Holders:** - Access to services **Local Communities:** - Access to Material Resources
Cluster 1	Issues on Employment and Decent Work	- Revenue - Cost - Net added Value	- Land Transformation - Water Footprint - Climate Impact	**Small Holders:** - Access to Services **Local Communities:** - Access to Material Resources - Local Employment - Secure Living Condition **Workers:** - Employment Relationship - Health and Safety - Fair Salary - Equal Opportunities - Working Hours **Value Chain Representatives:** - Fair Competition - Promoting Social Responsibility
Cluster 2	Future Technology and Education Requirements	- Revenue - Cost - Net added value	- Human Toxicity - Energy Types and Demands	**Small Holders:** - Access to services

3. social and environmental farming challenges.

Despite the differences in the clustering structure, the impact categories and subcategories identified as of high relevance and priority to stakeholders were largely consistent with those derived from k-means clustering.

These were land transformation, climate impact, water footprint, resource depletion, photochemical oxidation, local community access to material resources, employment qualities for workers, promotion of social responsibilities within the value chain, contribution to economic development for societies, and potential cost/revenue as indicators of financial performance. (Detailed results for the EM clustering can be found in Supplementary Material G).

4. Discussion

4.1. Comparison of manual and machine assisted approaches

Although, the manual approach did allow life cycle practitioners to assign impact categories based on stakeholder priorities, they are directly influenced by their involvement with stakeholders. Meanwhile, A computational linguistic expert is instead used separately to conduct the NLP and clustering process in the machine-assisted approach. This meant that the machine-assisted approach ensured that results could not be traced back to individual respondents by the life cycle practitioner, preserving stakeholder anonymity when processing stakeholder responses. Moreover, by reducing responses to their linguistic features and using heuristic based quantitative driven techniques to derive patterns, transparency and reproducibility of results from the machine assisted approach is enhanced when compared to the manual approach.

In terms of human resources, the manual approach required three LCA practitioners and one local translator while the machine-assisted approach required a minimum of three LCA practitioners, one computational linguistics expert, and one local translator. As machine learning and NLP methods continue to develop, they will likely allow for faster identification of high-priority impact categories and subcategories, even when dealing with complex arrays of stakeholder groups. This allows for a potentially more efficient scale up of the machine-assisted approach for larger life cycle study projects when compared to the manual alternative. The machine-assisted approach also enables more focus on themes of common interest shared across multiple stakeholder groups and are perceived to represent the interests and concerns of stakeholders as a whole, as opposed to the more individualised focus of the manual approach.

Although reducing part of the effort required to analyse data with the manual approach, some human involvement within the machine assisted approach is still necessary. Unlike typical classification models, which rely on predefined categories and performance metrics (e.g., precision, recall, false positives), the current method requires life cycle practitioners to manually interpret the thematic clusters. This helps ensure that the identified stakeholder concerns align with the study objectives rather than being dictated purely by statistical relationships within the data.

While the clustering step organises stakeholder responses into cohesive groups, human expertise is still required to contextualise and assign themes to these clusters in a way that reflects real-world implications. Moreover, due to the overlapping nature of certain stakeholder concerns, especially in EM clustering, experts must reach a consensus on appropriate labels. Thus, some level of human judgment remains necessary to validate and refine the assignment of impact categories.

4.2. Practical implications for life cycle studies

Taking into account the impact categories identified using the proposed machine-assisted approach not only allows the objectives of a life cycle study to be met but also helps guide resource allocation more efficiently, making sure data is collected, impacts are assessed, and conclusions are drawn on impacts of high relevance to stakeholders. Examples of how the themes in Tables 6 and 7 can be translated into impact categories that further guide decision making in the development of the life cycle inventory, impact assessment and interpretation phases of the life cycle study can be found in supplementary material H (Table-H).

One key example is the identification of future technology and education requirements using k-means clustering, which led to an increased emphasis on energy demand and human toxicity in the environmental impact assessment. This ensured that the energy consumption and emissions associated with new technologies were carefully considered during data collection and impact modelling. Similarly, EM clustering identified social and environmental challenges in farming,

Table 7
High priority impact categories and subcategories for each EM based cluster of responses and features.

Cluster	Cluster theme	LCC Impact Categories	Environmental LCA Impact Categories	Social LCA Impact Categories and Subcategories
Cluster 0	Economic Development Challenges	- Revenue - Cost	- Climate Impact - Water Footprint - Resource Depletion	**Workers:** - Health and Safety - Employment Relationships **Local Communities:** - Secure Living Conditions - Access to Material Resources **Society**: - Contribution to Economic Development
Cluster 1	land management funding and investments	- Cost	- Land Transformation - Water Footprint - biotic resource depletion	**Local Communities:** - Access to Material Resources - Local Employment **Workers:** - Equal Opportunities
Cluster 2	Social and environmental challenges in farming	- Revenue	- Photochemical Oxidation - Climate Impact - Resource Depletion	**Workers:** - Health and Safety - Equal Opportunities - Trading Relationships **Value Chain Representatives:** - promoting social responsibilities

highlighting climate impact, resource depletion, and employment conditions. This prompted a more focused evaluation of the social implications of farming practices in the SLCA, particularly regarding worker health, safety, and fair employment conditions. Additionally, clustering also emphasised economic development challenges, which informed the integration of revenue generation, cost structures, and financial feasibility, ensuring that these aspects were adequately integrated into the LCC.

However, prioritisation does not imply exclusion of other impact categories. Instead, it helps guide strategic allocation of time and data collection efforts to maximise the depth and relevance of analysis. LCSA aims to provide a comprehensive sustainability assessment while recognising practical constraints on data availability, time, and resources. The prioritisation process helps identify the most stakeholder-relevant categories for detailed examination and therefore potentially allowing the study to be methodologically rigorous and relevant for decision-making. At the same time, a broad set of categories is still considered to prevent burden shifting and to capture important sustainability trade-offs. By integrating machine-assisted techniques, practitioners can navigate these constraints more systematically. All relevant impacts are still acknowledged and contextualised in the interpretation phase, even those not selected for deeper assessment.

A major advantage of this data-driven stakeholder analysis is its ability to process large, unstructured datasets enabling practitioners to incorporate diverse perspectives while minimising individual biases. This can be particularly beneficial for multi-stakeholder projects, where diverse perspectives must be integrated into a single assessment. While this study focused on Namibia, the project itself was performed within had a wider scope, including Botswana and South Africa, where stakeholder priorities differ. The machine-assisted approach can help narrow down the diverse responses from these three countries to a manageable set of stakeholder preferences. Additionally, automating identification of stakeholder concerns and priorities improves reproducibility, so results should remain consistent across different operating practitioner teams.

The machine assisted goal and scope approach can also be expanded to projects involving stakeholders across multiple continents and languages. As technology advances, incorporating more African languages into translation platforms, POS tagging, and lemmatisation will make the digitalisation of stakeholder responses and preprocessing of text more automated. However, improvements in the efficiency of applying machine learning and NLP methods to stakeholder engagement will require further investigation. It is important to note that this study was achieved by using a relatively small dataset. As the size and complexity of datasets increase, the time-saving advantages of the machine-assisted approach will become even more pronounced.

Beyond life cycle studies, this approach has potential applications in sustainability reporting, environmental policy analysis and corporate social responsibility assessments. The ability to systematically extract and group stakeholder concerns efficiently could be valuable in sectors where stakeholder engagement is critical, such as renewable energy development, and circular economy initiatives.

4.3. Limitations and future research

While this study demonstrates the potential of NLP and ML for stakeholder response analysis, several limitations must be acknowledged, and future research should refine and expand upon the approach.

One limitation of this study is the lack of recorded demographic data (e.g., gender, age, education) for the stakeholder groups. As a result, potential biases related to representation and diversity in stakeholder perspectives could not be systematically assessed. While this study aimed to develop and test a machine-assisted approach for impact category selection rather than analyse demographic influences on stakeholder concerns, future research should integrate structured demographic data collection to allow for more detailed subgroup analyses. This would help assess whether specific demographic factors influence stakeholder priorities and how they may affect the relevance of different impact categories in life cycle studies.

Another challenge lies in extracting causal relationships between stakeholder concerns and impact categories. For example, biodiversity was identified as a priority in the manual approach but was not captured in the machine-assisted clustering results. This is due to the NLP model's reliance on surface-form word matching, which does not account for deeper contextual meaning or causality. In order to identify biodiversity-related features, observations must be conducted at the

elementary discourse unit level, leveraging lexico-syntactic patterns to extract causality. Without causality extraction, phrases like 'bush encroachment affects animal pathways' are reduced to individual words ('bush,' 'effects,' 'pathways', 'animal'), potentially leading to misclassification. When this text is presented to life cycle practitioners, these individual words may be overshadowed by other cluster features, shifting the interpretation from biodiversity impacts to resource depletion. When compared to other NLP-based text analysis approaches, such as those used in (Pechsiri et al., 2016), it becomes evident that integrating reasoning-based techniques, including textual causality mining (Pechsiri and Kawtrakul, 2007), causality extraction (Pechsiri and Piriyakul, 2010), causal web determination (Pechsiri et al., 2020), and deep learning for causality mining (Ali et al., 2021) — could significantly improve the machine-assisted goal and scoping process. Future studies should investigate the integration of these advanced NLP methodologies to enhance the ability of machine-assisted approaches to capture and correctly categorise causal relationships in stakeholder concerns.

A key limitation of the machine-assisted approach is the subjectivity involved in selecting the number of clusters, or k-value, which can significantly impact the analysis (Ikotun et al., 2023). Unlike typical classification methods, where predefined categories exist, unsupervised clustering requires determining an appropriate k-value, which is often based on expert judgment rather than an optimised statistical method. In this study, the k-value was pragmatically set to 3 to illustrate how clustering can be applied in the goal and scope phase of a life cycle study. The objective was not to capture all possible stakeholder concerns comprehensively but rather to demonstrate the feasibility of machine-assisted clustering for structuring stakeholder responses. Given that life cycle studies often face resource constraints that limit the extent of primary data collection, prioritising which impact categories to analyse in greater detail becomes essential. Setting k to 3 allowed for a manageable number of stakeholder concerns to be grouped into distinct themes, facilitating subsequent data collection and impact assessment.

While the chosen k-value was guided by domain knowledge, alternative methods exist for determining an optimal number of clusters (Naeem and Wumaier, 2018). Studies in other fields, such as healthcare, have set k-values based on expert knowledge of structured domains, such as diseases and major body systems (Pechsiri and Piriyakul, 2016b). Similarly, statistical approaches such as silhouette analysis, the elbow method, or stability-based validation could be applied to refine cluster selection. In this study, k-values ranging from 2 to 10 were tested (Supplementary Material D), but a systematic optimisation process was beyond the study's scope. Future research should explore approaches for optimising k-values in stakeholder response clustering, including the integration of factor analysis method as dimensionality reduction technique that focuses on identifying latent factors that explain correlations between variables or features [80, 81]. Additionally, identifying key performance metrics for validating clustering results in life cycle studies remains an important area for further investigation.

Scalability and language limitations present additional concerns. This study was conducted in Namibia, where stakeholder responses were translated and processed using English-based NLP tools. The effectiveness of this method in multilingual settings, particularly for languages with limited NLP resources, remains uncertain. Future research should explore how to adapt NLP models to underrepresented languages, especially in regions where life cycle studies require engagement with linguistically diverse communities. Advances in machine translation, part-of-speech tagging, and lemmatisation could improve the automation of stakeholder response processing, enabling more accurate clustering of key concerns across different linguistic contexts. As NLP technologies evolve, integrating African and other underrepresented languages into translation platforms and computational linguistic frameworks will be essential for expanding the applicability of machine-assisted stakeholder analysis in life cycle studies. However, further research is needed to enhance the efficiency and accuracy of unsupervised machine learning and NLP methods in these multilingual contexts.

Another limitation relates to the need for more automation in expert interpretation. Although the machine-assisted approach successfully automated the categorisation of stakeholder concerns, the final assignment of thematic labels and impact categories still required expert judgment. This introduces a degree of subjectivity and manual effort. Future studies should investigate semi-supervised or reinforcement learning techniques that allow models to learn from expert-labelled data, gradually improving their ability to assign themes and impact categories with reduced human intervention.

By addressing these limitations and research gaps, future studies can improve the robustness, scalability, and automation of machine-assisted goal and scope methods, further enhancing their contribution to life cycle studies and other decision-making frameworks.

5. Conclusion

The proposed machine-assisted approach significantly narrowed a broad range of impact categories to a more focused, prioritised selection. This study demonstrates the potential use of machine learning and natural language processing techniques during the goal and scope phase of a life cycle study, specifically for analysing stakeholder responses from engagement activities. The results demonstrate the successful application of both NLP and machine learning to process stakeholder responses.

However, several challenges remain. Optimising the number of clusters, incorporating languages that are not yet well supported by NLP models, improving causality extraction from responses, and addressing resource constraints are areas that require further research. While this method enhances objectivity, it does not eliminate the need for expert interpretation, particularly in refining thematic assignments and ensuring contextual accuracy.

One of the most important advantages of this method is its ability to maintain objectivity when analysing the data from stakeholder engagements. By emphasising specific impact categories based on stakeholder input rather than expert judgment, life cycle practitioners could address the most critical concerns raised by a diverse and complex network of stakeholders in Namibian bio-energy value chains with greater precision and neutrality.

By integrating NLP and ML into stakeholder response analysis, this study opens new possibilities for automating and streamlining stakeholder engagement in life cycle studies. Future advancements in NLP, clustering optimisation, and multilingual processing could further improve this approach, ensuring more transparent, efficient, and data-driven assessments.

CRediT authorship contribution statement

Joseph Santhi Pechsiri: Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alexandre Monteiro Souza:** Writing – original draft, Validation, Resources, Methodology, Investigation, Formal analysis. **Chaveevan Pechsiri:** Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis. **Paulus Kapundja Shigwedha:** Writing – original draft, Validation, Resources, Investigation. **Uasora Katjouanga:** Resources, Investigation. **Benjamin Mapani:** Writing – review & editing, Funding acquisition. **Rosa C. Goodman:** Writing – review & editing, Writing – original draft, Investigation. **Cecilia Sundberg:** Writing – review & editing, Investigation, Funding acquisition. **Niclas Ericsson:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition.

Compliance with Ethical standards

- Authors have no competing, no conflict, and no proprietary interests to declare.

- Authors have informed consent from all stakeholders. No experimentation of humans or animals involved.
- Funding is from the EU Horizon Programme

Data availability Statement

Response texts from stakeholders are on Supplement A, clustering preparation and experiment is in Supplement A and B. The resulted clusters are in Supplementary C and their assigned impact categories in D. See DOI repositories 10.5281/zenodo.13273288, 10.5281/zenodo.14000015, and 10.5281/zenodo.13735836 for info on SteamBioAfrica.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: All Authors reports financial support was provided by the European Union's Horizon 2020 research and innovation programme under grant agreement No 101036401.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101036401.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cesys.2025.100288.

Data availability

Data is shared as supplementary material. Project related data is provided as DOI links to repository. All Stakeholder engaged had signed consent forms. All information has been anonymized.

References

Abdella, G.M., Kucukvar, M., Onat, N.C., Al-Yafay, H.M., Bulak, M.E., 2020. Sustainability assessment and modeling based on supervised machine learning techniques: the case for food consumption. J. Clean. Prod. 251, 119661. https://doi.org/10.1016/j.jclepro.2019.119661.

Abokersh, M.H., Vallès, M., Cabeza, L.F., Boer, D., 2020. A framework for the optimal integration of solar assisted district heating in different urban sized communities: a robust machine learning approach incorporating global sensitivity analysis. Appl. Energy 267, 114903. https://doi.org/10.1016/j.apenergy.2020.114903.

Ali, W., Zuo, W., Ali, R., Zuo, X., Rahman, G., 2021. Causality mining in natural languages using machine and deep learning techniques: a survey. Appl. Sci. 11 (21), 10064.

Aloise, D., Deshpande, A., Hansen, P., Popat, P., 2009. NP-hardness of Euclidean sum-of-squares clustering. Mach. Learn. 75, 245–248.

Amutenya, F., 2021. Space-time Modelling of Unemployment Rate in Namibia.

Andreasi Bassi, S., et al., 2023. Updated characterisation and normalisation factors for the Environmental Footprint 3.1 method. JRC130796 Jt. Res. Cent.

Azari, R., Garshasbi, S., Amini, P., Rashed-Ali, H., Mohammadi, Y., 2016. Multi-objective optimization of building envelope design for life cycle environmental performance. Energy Build. 126, 524–534. https://doi.org/10.1016/j.enbuild.2016.05.054.

Bachmann, T.M., 2013. Towards life cycle sustainability assessment: drawing on the NEEDS project's total cost and multi-criteria decision analysis ranking methods. Int. J. Life Cycle Assess. 18, 1698–1709.

Bouckaert, R.R., et al., 2016. WEKA Manual for Version 3-9-1. Univ. Waikato Hamilt. N. Z., pp. 1–341

Brüntrup, M., Herrmann, R., 2012. 5. Bush-to-energy value chains in Namibia. In: Van Dijk, M.P., Trienekens, J. (Eds.), Global Value Chains. Amsterdam University Press, pp. 89–116. https://doi.org/10.1515/9789048514991-005.

Carbonell, J.G., Michalski, R.S., Mitchell, T.M., 1983. Machine learning: a historical and methodological analysis. AI Mag. 4 (3), 69–69.

Chiu, M.-C., Tai, P.-Y., Chu, C.-Y., 2024. Developing a smart green supplier risk assessment system integrating natural language processing and life cycle assessment based on AHP framework: an empirical study. Resour. Conserv. Recycl. 207, 107671. https://doi.org/10.1016/j.resconrec.2024.107671.

Chowdhury, G., 2003. Natural language processing. Annu. Rev. Inf. Sci. Technol. 37, 51–89.

Crenna, E., Secchi, M., Benini, L., Sala, S., 2019. Global environmental impacts: data sources and methodological choices for calculating normalization factors for LCA. Int. J. Life Cycle Assess. 24 (10), 1851–1877. https://doi.org/10.1007/s11367-019-01604-y.

Curran, M.A., 2016. Overview of goal and scope definition in life cycle assessment. In: *Goal and Scope Definition in Life Cycle Assessment*, in LCA Compendium – the Complete World of Life Cycle Assessment. Springer Dordrecht.

Damodaran, A., 2017. Narrative and Numbers: the Value of Stories in Business. Columbia Business School Publishing, Columbia University Press, New York.

Darena, F., Zizka, J., Burda, K., 2012. Grouping of customer opinions written in natural language using unsupervised machine learning. Presented at the 2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing. IEEE, pp. 265–270.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Methodol. 39 (1), 1–22.

Dogan, A., Birant, D., 2021. Machine learning and data mining in manufacturing. Expert Syst. Appl. 166, 114060.

Fellbaum, C., 2010. WordNet. In: Theory and Applications of Ontology: Computer Applications. Springer, pp. 231–243.

Geilfus, F., 2008. 80 Tools for Participatory Development: Appraisal, Planning, Follow-Up and Evaluation. San Jose: IICA.

Ghoroghi, A., Rezgui, Y., Petri, I., Beach, T., 2022. Advances in application of machine learning to life cycle assessment: a literature review. Int. J. Life Cycle Assess. 27 (3), 433–456. https://doi.org/10.1007/s11367-022-02030-3.

Guinee, J.B., et al. (Eds.), 2002. *Handbook on Life Cycle Assessment: Operational Guide to the ISO Standards*, Vol. 7. Eco-Efficiency in Industry and Science, vol. 7. Springer Netherlands, Dordrecht. https://doi.org/10.1007/0-306-48055-7.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. 11 (1), 10–18.

Hunkeler, D., Lichtenvort, K., Rebitzer, G., 2008. Environmental Life Cycle Costing. Crc press.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B., Heming, J., 2023. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data. Inf. Sci. 622, 178–210.

ISO, "Popular Standards ISO 14000 family," ISO - ISO 14000 family - Environmental Management. Accessed: March. 17, 2025. [Online]. Available: https://www.iso.org/standards/popular/iso-14000-family.

ISO, 2006a. Environmental Management - Life Cycle Assessment - Principles and Framework (ISO 14040:2006). European Committee for Standardization (CEN), Brussels. ISO-standard.

ISO, 2006b. ISO 14040: 2006 - Environmental Management - Life Cycle Assessment - Principles and Framework.

ISO, 2024. *ISO 14075:2024 Environmental management - Principles and framework for social life cycle assessment*, ISO 14075:2024(en). Geneva, Switzerland.

Jongejan, B., Dalianis, H., 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in-and suffixes alike. Presented at the Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 145–153.

Jongejan, B., Haltrup, D., 2005. The CST lemmatiser. Cent. Sprogteknologi Univ. Cph. Version 2.

Jung, Y.G., Kang, M.S., Heo, J., 2014. Clustering performance comparison using K-means and expectation maximization algorithms. Biotechnol. Biotechnol. Equip. 28 (Suppl. 1), S44–S48.

Kirill, M., Claudia, D.N., Andreas, C., Michael, S., 2020. PSILCA database v.3 documentation. GreenDelta [Online]. Available: https://psilca.net/wp-content/uploads/2020/06/PSILCA_documentation_v3.pdf. (Accessed 9 June 2024).

Klöpffer, W., 2008. Life cycle sustainability assessment of products: (with Comments by Helias A. Udo de Haes, p. 95). Int. J. Life Cycle Assess. 13, 89–95.

Krantz, A., Korn, R., Menninger, M., 2009. Rethinking museum visitors: using k-means cluster analysis to explore a museum's audience. Curator Mus. J. 52 (4), 363–374.

Meng, F., LaFleur, C., Wijesinghe, A., Colvin, J., 2019. Data-driven approach to fill in data gaps for life cycle inventory of dual fuel technology. Fuel (Guildf.) 246, 187–195.

Miller, G.A., 1995. WordNet: a lexical database for English. Commun. ACM 38 (11), 39–41.

Mitchell, T.M., 2006. The Discipline of Machine Learning, vol. 9. Carnegie Mellon University, School of Computer Science, Machine Learning.

Miyazaki, A.D., Taylor, K.A., 2008. Researcher interaction biases and business ethics research: respondent reactions to researcher characteristics. J. Bus. Ethics 81, 779–795.

Mlunga, L., Gschwender, F., 2015. Bush encroachment, de-bushing and energy production in Namibia. In: Perspectives on Energy Security and Renewable Energies

in Sub-saharan Africa, first ed. Macmillan Education Namibia, Windhoek, Namibia, p. 8 [Online]. Available: https://su-plus.strathmore.edu/bitstream/handle/11071/4458/PerspectivesonenergysecurityandrenewableenergiesinSub-SaharanAfrica.pdf?sequence=1#page=306.

Naeem, S., Wumaier, A., 2018. Study and implementing K-mean clustering algorithm on English text and techniques to find the optimal value of K. Int. J. Comput. Appl. 182 (31), 7–14.

Nguvenjengua, H., Undji, V., 2017. The impact of rural infrastructure shortages on poverty and income inequality in Namibia. Civ. Eng. Siviele Ingenieurswese 2017 (9), 54–54.

NPC, 2023. The Root Causes of Poverty. National Planning Commission of Namibia, Windhoek, Namibia.

O'Connor, T.G., Puttick, J.R., Hoffman, M.T., 2014. Bush encroachment in southern Africa: changes and causes. Afr. J. Range Forage Sci. 31 (2), 67–88. https://doi.org/10.2989/10220119.2014.939996.

Pechsiri, C., Kawtrakul, A., 2007. Mining causality from texts for question answering system. IEICE Trans. Info Syst. 90 (10), 1523–1533.

Pechsiri, C., Piriyakul, R., 2010. Explanation knowledge graph construction through causality extraction from texts. J. Comput. Sci. Technol. 25 (5), 1055–1070.

Pechsiri, C., Piriyakul, R., 2016a. Developing a Why–How Question Answering system on community web boards with a causality graph including procedural knowledge. Inf. Process. Agric. 3 (1), 36–53.

Pechsiri, C., Piriyakul, R., 2016b. Extraction of a group-pair relation: problem-solving relation from web-board documents. SpringerPlus 5, 1–25.

Pechsiri, C., Piriyakul, R., 2021. Causal pathway extraction from web-board documents. Appl. Sci. 11 (21), 10342.

Pechsiri, C., Moolwat, O., Piriyakul, R., 2016. Web board question answering system on problem-solving through problem clusters. Presented at the Knowledge, Information and Creativity Support Systems: Selected Papers from KICSS'2014-9th International Conference, Held in Limassol, Cyprus, on November 6-8, 2014. Springer, pp. 161–175.

Pechsiri, C., Keeratipranon, N., Piriyakul, I., 2020. Causal web determination from texts. J. Adv. Inf. Technol. 11 (2).

Romagnoli, F., et al., 2024. Furcellaria lumbricalis macroalgae cascade biorefinery: a life cycle assessment study in the Baltic Sea region. J. Clean. Prod. 478, 143861.

Romeiko, X.X., et al., 2024. A review of machine learning applications in life cycle assessment studies. Sci. Total Environ. 912, 168969. https://doi.org/10.1016/j.scitotenv.2023.168969.

Rosenbaum, R.K., 2016. Selection of impact categories, category indicators and characterization models in goal and scope definition. In: *Goal and Scope Definition in Life Cycle Assessment*, in LCA Compendium – the Complete World of Life Cycle Assessment. Springer, Dordrecht, pp. 63–122.

Sala, S., Farioli, F., Zamagni, A., 2013. Progress in sustainability science: lessons learnt from current methodologies for sustainability assessment: Part 1. Int. J. Life Cycle Assess. 18, 1653–1672.

Santorini, B., 1990. Part-of-speech Tagging Guidelines for the Penn Treebank Project.

Schön, T., 2009. An Explanation of the Expectation Maximization Algorithm.

Schreiber, J.B., Pekarik, A.J., 2014. Using latent class analysis versus K-means or hierarchical clustering to understand museum visitors. Curator Mus. J. 57 (1), 45–59.

Swarr, T.E., et al., 2011. Environmental life-cycle costing: a code of practice. Int. J. Life Cycle Assess. 16 (5), 389–391. https://doi.org/10.1007/s11367-011-0287-5.

Technical Committee ISO/TC 207, 2006. Environmental Management - Life Cycle Assessment - Requirements and Guidelines ISO 14044:2006.

The European Parliament and of the Council, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). p. 88. Accessed: March. 17, 2025. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.

UNEP, 2020. Guidelines for Social Life Cycle Assessment of Products and Organizations 2020. United Nations Environmenta Programme (UNEP) [Online]. Available: https://www.lifecycleinitiative.org/wp-content/uploads/2021/01/Guidelines-for-Social-Life-Cycle-Assessment-of-Products-and-Organizations-2020-22.1.21sml.pdf. (Accessed 11 January 2024).

UNEP, 2021. Methodological Sheets for Subcategories in Solcial Life Cycle Assessment (S-LCA) 2021. United nations Environment Programme (UNEP).

Wernet, G., Bauer, C., Steubing, B., Reinhard, J., Moreno-Ruiz, E., Weidema, B., 2016. The ecoinvent database version 3 (part I): overview and methodology. Int. J. Life Cycle Assess. 21 (9), 1218–1230. https://doi.org/10.1007/s11367-016-1087-8.

Yadav, J., Sharma, M., 2013. A review of K-mean algorithm. Int. J. Eng. Trends Technol. 4 (7), 2972–2976.

Yamamoto, Y.T., 2012. Values, objectivity and credibility of scientists in a contentious natural resource debate. Public Underst. Sci. 21 (1), 101–125.

Zhao, Y., Zhang, Q., Li, F.Y., 2019. Patterns and drivers of household carbon footprint of the herdsmen in the typical steppe region of inner Mongolia, China: a case study in Xilinhot City. J. Clean. Prod. 232, 408–416. https://doi.org/10.1016/j.jclepro.2019.05.351.