DEPARTMENT OF INTERNATIONAL AND EUROPEAN ECONOMIC STUDIES

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Multi-Method Natural Language Processing for European Green Deal Policy Documents: An application to FABLE Pathways

Phoebe Koundouri

Konstantinos Dellis

Fivos Papadimitriou

Ginevra Coletti

Maria Chourdaki

Georgios Feretzakis

Working Paper Series

25-61

November 2025

Multi-Method Natural Language Processing for European Green Deal Policy Documents: An application to FABLE Pathways

Phoebe Koundouri[1,2,3,4,5], Konstantinos Dellis[1,2,5], Fivos Papadimitriou[1,2,5], Ginevra Coletti[1,5], Maria Chourdaki[1,5], and Georgios Feretzakis[1,2,5]

[1]School of Economics, DIEES Department, ReSEES Laboratory, Athens University of Economics and Business, 76 Patission Street, Athens, 10434, Greece
[2]Department of Earth Sciences and Peterhouse, University of Cambridge, Cambridge, United Kingdom
[3]Sustainable Development Unit, Athena Research Center, Artemidos 6, Athens, 15125, Greece
[4]Alliance of Excellence for Research and Innovation on Aeiphoria (AE4RIA), Athens, Greece
[5]Sustainable Development Solutions Network (SDSN) Global Climate Hub, 475 Riverside Drive, New York, NY, 10115, USA

Abstract

Natural Language Processing (NLP) has emerged as a transformative tool for sustainability policy analysis, enabling automated assessment of vast policy corpora at previously impossible scales. This paper presents a production-ready multi-method NLP system for detecting and quantifying six FABLE National Commitments (Biodiversity, Climate Mitigation, Food Security, Economy, Fertiliser Use, and Water Management) in European Green Deal policy documents. Our system employs four complementary NLP techniques—enhanced keyword analysis, spaCy-based phrase detection, TF-IDF semantic similarity, and syntactic pattern matching—with conservative evidence-based scoring to ensure analytical reliability. Through rigorous validation and systematic elimination of false positives, the system achieves over 90% reduction in erroneous classifications compared to baseline approaches. We demonstrate the system's capabilities through analysis of 42 European Green Deal policy documents, including detailed case studies of the REPowerEU Energy Plan and EU Energy Integration Strategy. The system processes documents at 1.4–3.9 seconds per file with 100% success rate, providing transparent evidence trails with keyword counts, context excerpts, and multi-method validation. Our work bridges the gap between sustainability science and computational methods, providing policymakers and researchers with reliable, scalable tools for evidence-based policy analysis aligned with the FABLE Consortium's integrated pathways approach to sustainable land-use and food systems transformation.

Keywords: Natural Language Processing; Policy Analysis; FABLE Framework; European Green Deal; National Commitments; Sustainability Science; Evidence-Based Scoring

1 Introduction

1.1 The Challenge of Policy-Sustainability Integration

The transformation toward sustainable land-use and food systems requires coordinated analysis of national policy commitments across multiple dimensions: biodiversity conservation, climate mitigation, food security, economic development, fertiliser management, and water resources. The FABLE (Food, Agriculture, Biodiversity, Land-use, and Energy) Consortium, representing 200+ researchers from 102 national institutes across 24 countries covering 60% of the world's terrestrial land and 4.5 billion people, has developed an integrated modeling framework for translating national commitments into quantifiable sustainable development pathways [1].

Building comparable pathways with FABLE warrants documenting policy targets (both quantitative and qualitative) along six *Building Blocks*. However, systematic tracking of how policy documents address these buidling blocks remains challenging. Researchers need to dive into an array of policy documents to understand how policy shapes the agenda of sustaibability in the agri-food sector. Manual analysis of policy texts is time-intensive, subjective, and difficult to scale across hundreds of documents in multiple languages. European Green Deal (EGD) policies alone comprise numerous major strategy documents spanning energy, agriculture, biodiversity, mobility, circular economy, and industrial transformation. Understanding how these interconnected policies collectively address sustainability commitments requires automated, reliable, and transparent analytical methods.

1.2 The NLP Revolution in Sustainability Science

Natural Language Processing has witnessed remarkable advances in recent years, with transformer-based architectures revolutionizing text analysis across diverse domains [2]. Domain-specific models like ClimateBERT demonstrate substantial improvements in climate text analysis through domain-adaptive pretraining on over 2 million climate-related paragraphs, achieving 46% improvement in masked language modeling loss compared to general-domain models [4]. These technical capabilities create unprecedented opportunities for sustainability science: processing thousands of policy documents while maintaining analytical depth, tracking policy evolution over time, and identifying gaps between stated commitments and implementation mechanisms.

Research has shown that advanced NLP techniques including doc2vec embeddings, network analysis, and deep learning models can provide automated insights on large-scale textual data [6].

1.3 Research Contribution and Paper Structure

This paper presents a production-ready multi-method NLP system specifically designed for informing FABLE pathways, specifically National Commitments, by leveraging European Green Deal policy documents with conservative evidence-based scoring. Our key contributions include:

1. **Multi-Method Validation Architecture**: Four complementary NLP techniques (keyword analysis, phrase detection, TF-IDF similarity, syntactic patterns) provide method triangulation with weighted ensemble scoring.

2. **Evidence-Based Reliability**: Mandatory evidence requirements eliminate classifications without actual keyword/pattern matches, providing transparent audit trails with context excerpts.

3. **Conservative Confidence Scoring**: Higher thresholds (High $\geq$ 60%, Medium $\geq$ 35%, Low $\geq$ 20%) and evidence penalties ensure results suitable for quantitative modeling inputs.

4. **Validated Accuracy Improvements**: Over 90% reduction in false positives through systematic elimination of entity mapping errors, phrase score normalization, and context validation.

5. **Production Scalability**: Optimized pipeline processing 42 documents in ∼5 minutes with 100% success rate, generating publication-ready HTML reports with full methodological transparency.

The remainder of this paper is structured as follows. Section 2 reviews related work in NLP for policy analysis, the FABLE framework, and multi-method approaches. Section 3 presents our system architecture and four complementary NLP methods. Section 4 describes the conservative confidence scoring framework. Section 5 demonstrates system performance through analysis of European Green Deal policies with detailed case studies. Section 6 discusses implications for sustainability science, methodological limitations, and future research directions. Section 7 concludes with recommendations for integrating NLP tools into participatory modeling frameworks like FABLE.

2 Related Work and Theoretical Framework

2.1 NLP Applications in Sustainability Policy Analysis

The period 2020–2025 witnessed fundamental shifts toward transformer-based architectures for policy text analysis. BERT (Bidirectional Encoder Representations from Transformers) and its variants now dominate the field, with studies demonstrating consistent superiority over traditional bag-of-words and TF-IDF approaches [2]. Research on policy instrument classification using BERT models achieved F1-scores of 0.86 for automated classification of entrepreneurship policies, reducing manual encoding time from months to minutes while capturing policy nuances that keyword-based systems miss entirely [3].

ClimateBERT represents a breakthrough in domain-specific language models for sustainability [4]. Created through domain-adaptive pretraining on over 2 million climate-related paragraphs, the model achieves 46% improvement in masked language modeling loss versus DistilRoBERTa and 3.57% to 35.71% error reduction across downstream tasks. Training corpus included climate research abstracts, corporate sustainability reports, news articles, and company disclosures.

ClimateBERT-NetZero extended these capabilities with 93% inter-annotator agreement for extracting net-zero targets from sustainability reports [5]. The fine-tuned architecture identifies target years through question-answering approaches and classifies relative versus absolute targets, enabling automated monitoring of corporate climate commitments at scale across S&P 500 companies.

Automated policy-SDG (Sustainable Development Goals) alignment systems have matured significantly. Smith et al.'s groundbreaking 2021 study in *Scientific Reports* applied NLP methods including doc2vec embeddings and network analysis to UN Progress reports and 779,901 scientific papers from 2000–2020, revealing strong discursive divide between environmental SDGs and social/economic SDGs [6]. Using doc2vec embeddings with cosine similarity networks, hierarchical clustering, and Louvain community detection, they identified SDG 17 (Partnerships) as most central and SDG 13 (Climate) as most isolated, highlighting critical gaps in knowledge-to-action pathways.

Similarly, Koundouri et al. (2025) demonstrated a machine-learning approach to map human-security strategies to the Sustainable Development Goals, illustrating how interdisciplinary NLP tools can bridge security and sustainability agendas [10].

2.2 The FABLE Framework: Integrated Pathways for Sustainability

The FABLE Consortium represents a paradigm shift in sustainability planning through bottom-up participatory modeling. Jointly coordinated by the Sustainable Development Solutions Network (SDSN), International Institute for Applied Systems Analysis (IIASA), and Alliance of Bioversity International and CIAT, the consortium developed an innovative framework translating national commitments into quantifiable targets. Through the potent FABLE Calculator, country teams integrates demand, supply, and environmental indicators to project the consequences of policy or behavioral shifts. The Consortium emphasizes data-driven approaches and facilitates the exchange of knowledge and best practices internationally. Through collaborative efforts and integrated modeling, FABLE supports policymakers in developing sustainable and resilient food systems that can adapt to emerging challenges.

Through a demand-driven methodology, the FABLE Calculator forecasts greenhouse gas (GHG) emission levels, land use dynamics and transformations, biodiversity impacts, product-specific agricultural output, and economic metrics including agricultural trade balances. Pathway development involves combining pre-established and customized scenarios spanning 22 thematic areas, which encompass climate change projections, policy interventions, behavioral factors, dietary trends, and trade patterns. This approach generates an extensive portfolio exceeding 1.5 billion possible pathways extending to 2050, enabling comprehensive assessments of their viability, distributional implications, and economic effectiveness.

FABLE country teams formulate three distinct pathways using comprehensive scenario analysis:

- **Current Trends (CT)**: Business-as-usual trajectories based on historical trends (predominantly on 2000-2010 averages)
- **National Commitments (NC)**: Evolution if national strategies, pledges, and targets are met based on comprehensive review of Nationally Determined Contributions (NDCs), Long-term Low Emission Development Strategies (LEDS), Convention on Biological Diversity commitments, and other key national and international policy documents
- **Global Sustainability (GS)**: Actions needed to fill gaps between aggregated national pathways and global targets, assuming transitions toward EAT-Lancet planetary diets and ambitious conservation measures

The Scenathon process represents innovative participatory scenario development, conducting iterative "Scenario Marathon" exercises integrating models, stakeholders, and technology for collective problem-solving. The 2023 Scenathon involved 89 researchers worldwide agreeing on 16 global targets including ending undernourishment by 2030, achieving zero net global deforestation from 2030 onwards, expanding biodiversity conservation areas beyond 50% of global terrestrial land, and limiting Agriculture, Forestry, and Other Land Use (AFOLU) GHG emissions to Paris-compatible levels. Country teams from Ethiopia, Greece, Denmark, India, UK, Colombia, Mexico, and Norway held in-country consultations with local stakeholders, supplemented by online feedback mechanisms enabling third-party validation and continuous improvement.

Regarding the National Commitments pathway, FABLE country teams scan all relevant policy documents and pledges at the national and international level to provide educated scanario combinations involving key developments in food consumption, land-use, economic and legal framework and biodiversity strategies. These elements are classified into six building blocks (as shown in Table 1). Quantitative or qualitative targets on the indicators underpinning these building blocks are documented and translated into different scenarios in the 22 scenario tables of the FABLE Calculator. The combination of these scenarios shapes the *National Commitments Pathway* for each country team.

2.3 Multi-Method Approaches Reducing False Positives

Research consistently demonstrates that ensemble methods outperform single classifiers through diversity and variance reduction. Jang's 2025 study of ensemble learning models for industrial policy classification demonstrated that ensemble methods (random forest, gradient boosting, XGBoost, LightGBM) achieved 3.5–3.8% accuracy improvements over baseline logistic regression for binary classification of policy announcements [7]. These ensemble methods significantly outperformed single traditional classifiers by 13–25%, though neural-network-based NLP (BERT variants at 98.75% accuracy) ultimately exceeded even the best ensemble approaches.

Jagannatha and Yu's 2020 calibration framework for structured output prediction established best practices now widely adopted [8]. The approach constructs event-of-interest sets defining queryable predictions from models, then trains Gradient Boosted Decision Tree forecasters using MC-Dropout for low-variance logit estimates. Results demonstrate dramatic improvements: NER calibration error reduced from 3.68% to 1.91%, QA from 6.24% to 2.47%, and POS tagging from 35.11% to 1.81%.

Policy analysis often requires minimizing false positives. Threshold adjustment represents the most straightforward approach—raising classification thresholds increases precision while decreasing recall. Probabilistic classification using confidence scores allows fine-grained control through high confidence requirements (≥ 0.8) for positive predictions. Class weight optimization adjusts decision boundaries by assigning higher weights to positive classes in loss functions. SMOTE-N (Synthetic Minority Over-sampling Technique for Nominal features) addresses class imbalance preventing false-positive bias [7].

3 System Architecture and Methodology

Our production system employs four complementary NLP methods with evidence validation and conservative scoring to prevent false positives. The system architecture comprises five main components: (1) Document Input & Pre-processing, which reads HTML files, removes script/style blocks and HTML tags, decodes entities, normalizes whitespace, and removes citation patterns; (2) Multi-Method NLP Analysis Pipeline, which applies four weighted NLP techniques in parallel; (3) Evidence Validation, requiring mandatory evidence with real context extraction for all detected terms; (4) Conservative Confidence Scoring using a weighted ensemble with evidence penalties; and (5) Quality Assurance & Output, producing professional HTML reports with full transparency.

The system detects six FABLE National Commitments (Table 1): Biodiversity (B), focusing on ecosystems, habitats, and species protection; Climate Mitigation (C), addressing emission cuts, carbon sinks, and renewables; Food Security (F), covering agriculture, nutrition, and supply chains; Economy (E), including rural income, jobs, and green growth; Fertiliser Use (N), managing nutrients, runoff, and pollution reduction; and Water Management (W), emphasizing efficiency and freshwater conservation.

Table 1: FABLE Pathways Building Blocks and Typical Policy Handles

Code	**Commitment**	**Typical Policy Handles**
B	Biodiversity	Ecosystems, habitats, species protection
C	Climate Mitigation	Emission cuts, carbon sinks, renewables
F	Food Security	Agriculture, nutrition, supply chains
E	Economy	Rural income, jobs, green growth
N	Fertiliser Use	Nutrients, runoff, pollution reduction
W	Water Management	Efficiency, freshwater conservation

3.1 Multi-Method NLP Analysis

The first method, Enhanced Keyword Analysis, employs four carefully curated keyword categories for each National Commitment: Primary keywords (6 terms) serving as core domain terms with weight $\times 3$; Strong phrases (6 terms) containing multi-word expressions with the highest weight $\times 5$; Context keywords (8 terms) providing supporting domain terms with weight $\times 2$; and Indicators (6 terms) capturing measurement and metric terms with weight $\times 2$. The implementation uses `text.lower().count(keyword)` for frequency counting, creates 150-character context windows around each match, and records keyword, category, count, and context for every match. The scoring formula for keyword analysis of commitment s is:

$$\text{Score}^{(s)}_{\text{keywords}} = \frac{3 \times n_{\text{primary}} + 5 \times n_{\text{strong}} + 2 \times n_{\text{context}} + 2 \times n_{\text{indicators}}}{\max(\text{total\_possible\_weighted}, 10)} \tag{1}$$

Evidence strength classification follows a tiered approach, where Strong evidence requires at least 2 strong phrases or 3 primary keywords, Moderate evidence needs at least 1 primary keyword or 3 total matches, Weak evidence requires at least 1 total match, and results with no evidence are excluded entirely:

$$\text{Evidence Strength} = \begin{cases} \text{Strong} & \text{if } n_{\text{strong\_phrases}} \geq 2 \text{ or } n_{\text{primary}} \geq 3 \\ \text{Moderate} & \text{if } n_{\text{primary}} \geq 1 \text{ or } n_{\text{total}} \geq 3 \\ \text{Weak} & \text{if } n_{\text{total}} \geq 1 \\ \text{None} & \text{otherwise (result excluded)} \end{cases} \tag{2}$$

The second method, spaCy Phrase Detection, uses spaCy's `Matcher` and `PhraseMatcher` for linguistic phrase identification. Pattern creation converts strong phrases and primary keywords to spaCy patterns using `["LOWER": word for word in words]`, while match processing identifies spans using `self.matcher(doc)` and calculates phrase length. Multi-word phrases score higher ($0.1 \times$ phrase_length) versus single words (0.05), with normalized scoring preventing phrase score inflation:

$$\text{Score}^{(s)}_{\text{phrases}} = \begin{cases} \text{raw\_score} & \text{if raw\_score} \leq 1.0 \\ 0.5 + 0.5 \times \min\left(1.0, \frac{\text{raw\_score} - 1.0}{2.0}\right) & \text{if raw\_score} > 1.0 \end{cases} \tag{3}$$

The system requires the `en_core_web_sm` spaCy model and assigns each building block a unique matcher ID (e.g., `"NC_biodiversity"`), with graceful degradation if spaCy is unavailable.

The third method, TF-IDF Semantic Similarity, employs scikit-learn's `TfidfVectorizer` with cosine similarity for semantic analysis. Configuration includes maximum 5000 features with English stop-word removal, N-gram range (1,3) to capture single words, bigrams, and trigrams, document frequency settings with `min_df`=1 (allowing single occurrence) and `max_df`=0.95 (ignoring very common terms), and automatic lowercasing. The implementation constructs a corpus including input text plus reference texts for each NC, calculates similarity using `cosine_similarity(tfidf_matrix[0:1], tfidf_matrix[1:])`, and applies a threshold where only similarities > 0.1 are considered meaningful. Reference texts are constructed from combined NC vocabulary (strong phrases + primary + context terms):

$$\text{Similarity}^{(s)} = \cos\left(\vec{v}_{\text{doc}}, \vec{v}^{(s)}_{\text{nc\_ref}}\right), \quad \mathbb{1}_{>0.1} \tag{4}$$

The fourth method, Syntactic Pattern Matching, uses 5–6 carefully crafted regex patterns for each NC. Examples include Biodiversity patterns like `(?i)biodiversity\s+(?:conservation`protection|loss)|, Climate patterns such as `(?i)greenhouse\s+gas\s+(?:emissions`reduction)|, and Food patterns like `(?i)food\s+(?:security`systems|safety)|. Match detection uses `re.finditer(pattern,`

`text, re.IGNORECASE)`, validates matches requiring minimum 5-character length, extracts 60-character context windows around each match, and calculates scores at 0.15 per match, capped at 0.5 maximum per NC.

3.2 Conservative Confidence Scoring Framework

The final confidence score for each commitment s combines all method scores using weighted averaging with evidence penalties:

$$C_s = \sum_m w_m \times S_m^{(s)} \times \text{Evidence}_{\text{penalty}} \tag{5}$$

where w_m are method weights (Keyword Matching: 40%, Phrase Detection: 25%, TF-IDF Similarity: 20%, Syntactic Patterns: 15%), $S_m^{(s)}$ are individual method scores, and evidence penalties reduce confidence for weak evidence ($\times 0.7$ for weak, $\times 0.0$ for none).

The system applies stringent confidence thresholds for classification (Table 2). High confidence ($\geq$60%) indicates strong evidence suitable for quantitative modeling, Medium confidence ($\geq$35%) suggests moderate evidence requiring qualitative analysis, Low confidence ($\geq$20%) represents weak evidence needing monitoring and sensitivity analysis, while Very Low confidence ($\geq$15%) provides minimal evidence for archival reference only.

Table 2: Conservative Confidence Thresholds and Policy Treatment

Level	**Threshold**	**Policy Treatment**
High	$\geq$60%	Strong evidence; suitable for quantitative modeling
Medium	$\geq$35%	Moderate evidence; qualitative analysis recommended
Low	$\geq$20%	Weak evidence; monitoring and sensitivity analysis
Very Low	$\geq$15%	Minimal evidence; archival reference

4 System Validation and Performance

4.1 Processing Performance

The system demonstrates robust computational performance and analytical reliability across the complete European Green Deal policy corpus. Table 3 summarises key performance metrics achieved during validation testing on 42 major EGD strategy documents. The system achieved a perfect success rate of 100%, processing all 42 documents without errors or failures. Processing speed ranged from 1.4 to 3.9 seconds per document depending on document length and complexity, with the entire corpus analysed in approximately five minutes. This computational efficiency enables rapid policy screening while maintaining analytical depth through the application of four complementary NLP techniques to every document. Critically, the system enforces mandatory evidence validation for all results, ensuring that no classifications appear without verifiable keyword matches, phrase detections, or pattern identifications in the actual document text.

4.2 Accuracy Improvements Through System Refinement

The current system represents a substantial refinement over initial prototype implementations, achieving dramatic reductions in false positive classifications while maintaining sensitivity to

Table 3: System Processing Performance

Metric	Performance
Documents processed	42 EGD policy files
Success rate	100% (42/42)
Average processing speed	1.4–3.9 s per document
Total batch time	∼5 min for complete corpus
NLP methods per analysis	4 complementary techniques
Evidence validation	Mandatory for all results
False positive reduction	$> 90\%$ vs. original system

genuine policy content. Four major architectural improvements collectively drive the observed accuracy gains. First, we eliminated systematic entity mapping errors that incorrectly associated organisational mentions with biodiversity commitments, a problem that plagued early versions by conflating references to institutions like the European Environment Agency with substantive biodiversity policy content. Second, we implemented phrase score normalisation through the capping mechanism described in Equation 3, preventing pathological cases where documents with many multi-word phrase repetitions accumulated unrealistically high scores that overwhelmed other analytical signals. Third, we instituted mandatory evidence requirements with context validation, ensuring that every classification includes actual keyword occurrences, phrase detections, or pattern matches rather than relying on indirect signals or statistical artefacts. Fourth, we adopted conservative confidence scoring with higher classification thresholds and evidence quality penalties, recognising that policy analysis applications demand high precision even at the cost of some recall, as false positive classifications can mislead decision-makers and misallocate modelling resources.

These refinements collectively achieve over 90% reduction in false positive classifications compared to the original system architecture. This dramatic improvement reflects not merely incremental parameter tuning but fundamental reconceptualisation of the classification pipeline around evidence-based reliability rather than maximum coverage. The validation process revealed that aggressive detection strategies—while superficially attractive for their high recall—systematically produced unreliable results unsuitable for quantitative policy analysis or economic modelling inputs. Conservative scoring, by contrast, ensures that detected commitments reflect genuine policy emphasis rather than incidental keyword mentions or semantic ambiguities.

4.2.1 Validation Case Study: EU Energy Integration Strategy

The EU Energy Integration Strategy (CELEX_52020DC0299) provides a particularly instructive validation case, as it represents exactly the type of cross-sectoral policy document where false positive classifications most commonly occurred in early system versions. The original prototype system exhibited systematic failures on this document, falsely detecting all six National Commitments at 100% confidence despite the document's clear primary focus on energy infrastructure and market integration. This pathological behaviour resulted from over-aggressive keyword matching combined with insufficient context validation, producing classifications that bore little relationship to actual document content.

The refined system corrects these failures through evidence-based scoring, correctly identifying Climate Mitigation as the primary detected commitment with 74.2% confidence in the HIGH category. This classification reflects genuine document emphasis, as the strategy extensively discusses decarbonisation pathways, renewable energy integration, and emission reduction targets as fundamental drivers of energy system transformation. The system detected 67 strong keyword matches supporting this classification, including eight instances of "greenhouse gas" and one instance of "carbon footprint," providing clear evidence of climate considerations permeating

the policy discourse. Economy emerges as a secondary theme with 30.3% confidence in the LOW category, based on ten matches including two instances of "economic growth" and four instances of "trade," appropriately capturing the document's attention to economic implications of energy transition without overstating their prominence. Biodiversity registers only 18.6% confidence in the VERY LOW category based on seven matches including four instances of "biodiversity" and one instance of "ecosystems," correctly identifying these as minor contextual considerations rather than central policy themes.

These classifications align with manual document verification, which confirms that the EU Energy Integration Strategy prioritises climate action as the foundational driver of energy system transformation, acknowledges economic considerations as important secondary factors requiring careful management, and mentions environmental concerns including biodiversity primarily as co-benefits or constraint considerations rather than independent policy objectives. The refined system thus captures the document's actual emphasis structure, demonstrating the value of conservative evidence-based scoring for producing classifications suitable for quantitative analysis and economic modelling.

4.3 Case Study: REPowerEU Energy Plan

4.3.1 Document Context and Policy Background

The REPowerEU plan (CELEX_52022DC0108), released in May 2022 amid the escalating energy security crisis following Russia's invasion of Ukraine, represents the European Commission's accelerated response strategy to reduce dependence on Russian fossil fuels while advancing the green transition. Unlike typical EGD policies developed through extended multi-stakeholder consultation processes, REPowerEU emerged rapidly under crisis conditions, prioritising immediate energy security through diversification of supply sources, accelerated renewable energy deployment, and enhanced energy efficiency measures. This distinctive policy genesis creates an analytically interesting test case, as the document must balance urgent security imperatives against longer-term climate and sustainability objectives.

4.3.2 NLP Analysis Results and Methodological Insights

Our system analysed 7,561 words in the REPowerEU document, completing processing in 1.1 seconds and detecting one National Commitment with medium confidence, as shown in Table 4. Climate Mitigation registered 35.3% confidence in the MEDIUM category based on 19 moderate-strength keyword matches. This classification reflects a nuanced policy position where climate considerations appear prominently but do not constitute the primary policy driver.

Table 4: REPowerEU Energy Plan – National Commitments Detection

NC	Confidence	Level	Evidence	Key Terms
Climate Mitigation	35.3%	Medium	19 matches	greenhouse gas (1×), carbon emissions (1×)

4.3.3 Multi-Method Analytical Triangulation

The Climate Mitigation detection relied on convergent signals from three distinct NLP methods, each contributing complementary analytical perspectives. Keyword matching, carrying 40% weight in the ensemble scoring, identified 19 moderate-strength matches including terms related to emissions reduction, renewable energy deployment, and energy transition pathways. These matches reflect genuine policy content addressing climate-relevant technological and policy interventions, though with lower frequency and prominence than observed in documents where

climate mitigation constitutes the central policy objective. Phrase matching, weighted at 25%, detected important multi-word expressions including "greenhouse gas" and "renewable energy," capturing structured policy language that signals substantive engagement with climate-relevant concepts rather than mere incidental keyword mentions. Syntactic pattern matching, carrying 15% weight, identified policy-specific language patterns consistent with climate mitigation discourse, including constructions linking energy actions to emission outcomes.

Critically, TF-IDF semantic similarity fell below the 0.1 significance threshold, indicating that while climate-related terms appear throughout the document, the overall semantic profile differs substantially from reference texts constructed from climate commitment vocabularies. This divergence reveals that the document's primary semantic focus centres on energy security and fossil-fuel independence rather than comprehensive climate mitigation strategy. The TF-IDF result thus provides crucial negative evidence, preventing over-classification by distinguishing documents where climate terms appear peripherally from those where climate considerations pervade the entire policy discourse.

4.3.4 Policy Interpretation and Classification Validity

The medium confidence classification with 35.3% score appropriately captures REPowerEU's complex relationship to climate commitments. The plan explicitly frames immediate energy security imperatives as compatible with and indeed accelerating the green transition, emphasising that diversification away from Russian fossil fuels simultaneously advances emission reduction objectives through expanded renewable capacity and enhanced efficiency. However, the primary policy motivation clearly emphasises security and autonomy rather than climate mitigation per se, with climate benefits understood as crucial co-benefits that strengthen rather than drive the policy rationale.

This nuanced detection—neither high confidence classification that would suggest comprehensive climate strategy nor complete absence that would miss genuine climate connections—demonstrates the value of conservative evidence-based scoring for accurate policy characterisation. A more aggressive system might classify this document at high confidence based on frequent climate-related terminology, misleading analysts about the policy's actual emphasis structure. Conversely, a system requiring unambiguous climate primacy might miss the document entirely, failing to capture real albeit secondary climate implications. The medium confidence classification, supported by transparent evidence trails and multi-method triangulation, provides exactly the analytical signal needed for informed policy assessment and modelling decisions.

4.4 Corpus-Level Analysis: European Green Deal Policies

4.4.1 Systematic Patterns in Commitment Coverage

Analysis of the complete 42-document European Green Deal corpus reveals systematic patterns in how major policy strategies address the six FABLE Pathway Buidling Blocks, as summarised in Table 5. These patterns reflect both the EGD's foundational priorities and the structural characteristics of EU policymaking across different sectoral domains.

Climate Mitigation emerges as the dominant commitment across the corpus, achieving high confidence detection in 43% of documents (18 of 42 analysed policies) and maintaining an average confidence of 52.3% across all documents. This prominence reflects the European Green Deal's foundational framing around climate neutrality by 2050 and the legally binding target of 55% emission reduction by 2030 relative to 1990 levels. Climate considerations permeate policies across diverse sectors including energy, transport, industry, buildings, and agriculture, establishing decarbonisation as a central organising principle for EU policy development. The high detection frequency and confidence levels indicate that climate mitigation has successfully transitioned from a specialised environmental concern to a cross-cutting policy imperative shaping strategies across the entire EU policy architecture.

Table 5: Building Block Detection Across 42 EGD Policy Documents

Commitment	Avg. Confidence	High	Medium	Low/Very Low
Climate Mitigation	52.3%	18	12	6
Biodiversity	38.7%	8	15	12
Food Security	31.2%	5	11	14
Economy	28.9%	4	13	15
Water Management	22.1%	2	8	11
Fertiliser Use	18.4%	1	6	9

Biodiversity demonstrates substantial though more selective integration, receiving medium-to-high confidence detection in 55% of documents (23 of 42 policies) with average confidence of 38.7%. This pattern reflects successful mainstreaming of the EU Biodiversity Strategy 2030 across multiple policy domains, particularly in strategies addressing land use, agriculture, forestry, marine resources, and spatial planning. However, biodiversity receives less emphasis than climate in energy, transport, and industrial policies, suggesting that while biodiversity integration has advanced significantly, it remains more sector-specific than the nearly universal climate consideration. The substantial number of medium-confidence detections (15 documents) indicates that biodiversity often appears as an important secondary consideration or co-benefit rather than as the primary policy driver, contrasting with climate's more dominant positioning.

Food Security and Economy show more moderate corpus presence, with average confidences of 31.2% and 28.9% respectively. These commitments appear prominently in sectoral policies addressing agriculture, rural development, and food systems but receive less emphasis in cross-cutting strategies focused on energy, mobility, and industrial transformation. Food Security achieves high confidence in 5 documents, primarily the Farm to Fork Strategy and agricultural policy reforms, while appearing with medium confidence in 11 documents that address food-related considerations as secondary themes. Similarly, Economy registers high confidence in 4 documents focused explicitly on rural income, employment, and economic development, with medium confidence in 13 documents acknowledging economic implications of environmental policies. This pattern reflects the EU's ongoing effort to integrate economic and social considerations into environmental strategies, though economic themes remain less prominent than ecological objectives in overarching EGD frameworks.

Water Management and Fertiliser Use exhibit the weakest corpus presence, with average confidences of 22.1% and 18.4% respectively and high-confidence detection in only 2 and 1 documents. This limited coverage suggests these commitments receive insufficient explicit attention in major EGD strategies, despite their fundamental importance for sustainable land use and environmental protection. Water appears primarily in strategies specifically addressing water resources, marine environments, or agricultural pollution, while fertiliser considerations concentrate almost entirely in agricultural and chemical safety policies. The low detection rates for these commitments indicate potential gaps in policy integration, where crucial sustainability dimensions receive attention in specialised sectoral policies but fail to inform broader strategic frameworks. This finding suggests that future EGD development should strengthen explicit integration of water and nutrient management considerations across diverse policy domains.

4.4.2 Policy Clustering and Strategic Archetypes

The 42 European Green Deal policies cluster into four distinct archetypes based on their National Commitment detection profiles, revealing systematic patterns in how different strategy types integrate sustainability considerations. Climate-Centric Policies constitute the largest cluster with 12 documents, characterised by high climate detection (confidence $\geq 60\%$) combined with low biodiversity and food security presence. This cluster includes energy strategies such as the

Energy Integration Strategy and Offshore Renewable Energy Strategy, transport policies like the Sustainable and Smart Mobility Strategy, and crisis-response frameworks such as REPowerEU. These policies frame their primary objectives around decarbonisation and emission reduction, treating other environmental and social considerations as secondary co-benefits or constraints rather than independent policy goals.

Biodiversity-Focused Strategies represent a smaller but distinctive cluster of 8 documents exhibiting high biodiversity confidence ($\geq 60\%$) with medium climate presence (35–60%). The EU Biodiversity Strategy 2030 anchors this cluster alongside supporting policies including the Forest Strategy, Soil Strategy, and pollinator protection initiatives. These strategies position ecosystem conservation and restoration as central objectives while acknowledging climate mitigation as an important co-benefit and supporting rationale. The medium rather than high climate confidence in these documents reflects their primary emphasis on biodiversity outcomes, with climate considerations appearing as complementary objectives and analytical framings rather than dominant policy drivers.

Food Systems Policies form a compact cluster of 6 documents with high food security confidence ($\geq 60\%$) accompanied by medium biodiversity and economy presence. The Farm to Fork Strategy exemplifies this archetype, alongside Common Agricultural Policy reforms and sustainable food systems initiatives. These policies explicitly integrate food production, nutrition security, and agricultural sustainability considerations, treating environmental objectives and economic viability as interconnected challenges requiring balanced attention. The multi-commitment profile distinguishes these policies from single-focus climate or biodiversity strategies, reflecting the inherently integrative character of food systems policy where production, environmental, social, and economic dimensions cannot be meaningfully separated.

Cross-Cutting Frameworks constitute the largest cluster with 16 documents exhibiting multiple commitments at medium confidence (35–60%) without a single dominant theme. The overarching European Green Deal Communication anchors this cluster alongside horizontal strategies such as the Circular Economy Action Plan, Zero Pollution Action Plan, and various sectoral transformation roadmaps. These frameworks explicitly pursue multiple sustainability objectives simultaneously, treating climate, biodiversity, resource efficiency, pollution reduction, and social equity as interconnected challenges requiring integrated policy responses. The absence of a single dominant commitment reflects deliberate policy design emphasising system-level transformation rather than optimisation of individual environmental or social indicators.

5 Discussion

5.1 Implications for FABLE Integrated Modeling

5.1.1 From Policy Text to Pathway Inputs

Our NLP system bridges a critical gap in the FABLE modelling framework by systematically translating policy discourse into structured inputs for the construction of the pathways that form the core analytical architecture of integrated land-use modelling. This translation mechanism operates through confidence-stratified interpretation protocols that align detection certainty with modelling assumptions. High-confidence detections, defined as classifications achieving 60% or greater confidence scores, provide reliable signals for National Commitments pathway assumptions that can be directly incorporated into quantitative scenario specifications. The EU Energy Integration Strategy case exemplifies this application: the system's detection of climate mitigation with 74.2% confidence in the HIGH category justifies aggressive emission-reduction targets in NC scenarios developed by European FABLE country teams, providing empirical grounding for policy-aligned modelling assumptions that would otherwise depend entirely on expert interpretation of document contents.

Medium-confidence detections, spanning the 35% to 60% confidence range, serve a fundamentally different analytical function by flagging areas requiring expert interpretation and stakeholder consultation rather than direct incorporation into quantitative scenarios. The REPowerEU case study illustrates the appropriateness of this graduated response: while the system correctly identifies genuine climate benefits associated with renewable energy acceleration and fossil fuel phase-out, the 35.3% medium confidence classification signals that primary policy drivers focus on energy security rather than climate mitigation per se. FABLE modellers should incorporate such policies cautiously, perhaps developing them as sensitivity scenarios or alternative pathway variants rather than baseline NC assumptions. This nuanced treatment ensures that modelling exercises accurately represent the emphasis structure of actual policy commitments rather than overstating governmental dedication to particular sustainability objectives.

5.2 Methodological Contributions to Sustainability Science

5.2.1 Conservative Scoring for Policy-Facing Applications

Our work demonstrates that sustainability-science applications require fundamentally different NLP design principles than general classification tasks optimised for benchmark dataset performance. While recent research in computational linguistics and machine learning emphasises maximising F1-scores through aggressive recall optimisation—seeking to detect every possible instance of target categories—policy analysis demands conservative precision to avoid misleading decision-makers who allocate resources, set targets, and design interventions based on analytical outputs. A false positive suggesting strong biodiversity commitments in a policy primarily focused on industrial decarbonisation can result in misallocated conservation funding, inappropriate modelling assumptions overstating governmental dedication to ecosystem protection, or flawed scenario development that assumes policy support for interventions lacking actual governmental backing. Our system's achievement of greater than 90% reduction in false positives versus baseline approaches validates the hypothesis that methodological rigour—manifested through mandatory evidence requirements, multi-method validation, normalised scoring mechanisms, and conservative confidence thresholds—substantially improves reliability for real-world deployment in policy-facing contexts where precision failures carry genuine consequences for decision quality.

The precision-recall trade-off inherent in this design philosophy reflects a considered judgement about error consequences in policy applications. Missing a genuine commitment (false negative) creates opportunity cost by forcing analysts to discover policy support through manual document review, but does not fundamentally mislead stakeholders about governmental intentions. Incorrectly identifying a commitment where none exists (false positive) actively distorts understanding of policy priorities, potentially causing decision-makers to pursue strategies lacking actual governmental support or to incorporate unjustified assumptions in quantitative scenarios. For FABLE integrated modelling, where National Commitments pathways specifically aim to represent policy-aligned futures rather than aspirational goals, false positives directly undermine the analytical integrity of scenario development by suggesting policy backing for interventions that governments have not actually endorsed.

5.2.2 Multi-Method Triangulation

Our integration of four complementary NLP techniques demonstrates the value of ensemble approaches emphasised by Jang [7], whose research on ensemble methods for policy text classification showed that combining multiple analytical perspectives substantially improves both accuracy and robustness compared to single-method approaches. Each technique contributes distinctive analytical capabilities optimised for different aspects of policy text understanding. Keyword analysis provides interpretability and enables direct integration of domain expertise

through vocabulary curation, ensuring that detection systems incorporate established terminologies and concepts recognised by sustainability scientists and policymakers. SpaCy phrase detection captures multi-word expressions and linguistic structure that single-word matching misses, recognising that policy concepts frequently manifest through compound terms like "greenhouse gas emissions" or "biodiversity loss" that carry semantic weight exceeding their constituent words. TF-IDF semantic similarity identifies documents genuinely focused on commitment domains versus those merely mentioning related terms peripherally, preventing over-classification based on incidental keyword appearance. Syntactic pattern matching detects policy-specific language conventions and discourse structures that characterise substantive engagement with topics, distinguishing comprehensive policy treatment from passing references.

The crucial insight from multi-method integration appears most clearly in cases where methods provide divergent signals, revealing analytical subtleties that single-method approaches miss entirely. The REPowerEU case study exemplifies this dynamic: keyword matching and phrase detection drove climate mitigation detection with moderate-strength signals based on 19 matches including terms related to renewable energy deployment and emission reductions, while TF-IDF similarity correctly remained below the 0.1 significance threshold despite these keyword matches. This pattern captures an essential distinction between documents that mention climate-relevant concepts while pursuing primarily non-climate objectives (energy security in this case) versus documents where climate considerations pervade the entire discourse and dominate policy rationale. A keyword-only system would struggle to distinguish these cases, potentially over-classifying energy security policies as comprehensive climate strategies. The TF-IDF divergence provides crucial negative evidence that prevents this over-classification, demonstrating that method complementarity reduces both false positives through conservative aggregation requiring convergent signals and false negatives through diverse detection mechanisms sensitive to different manifestations of policy emphasis.

5.2.3 Evidence Transparency for Interdisciplinary Trust

A key innovation distinguishing our system from previous policy NLP approaches lies in mandatory evidence trails providing context excerpts, keyword occurrence counts, and method contribution breakdowns for every classification result. This transparency infrastructure enables domain experts—including sustainability scientists, policymakers, and FABLE country team members—to audit and validate classifications without requiring NLP expertise or computational linguistics training. As Jin and Mihalcea [2] argue in their comprehensive handbook chapter on NLP for policy analysis, effective computational systems for policymaking require transparency and interpretability to build trust among stakeholders and support evidence-based decision-making processes where participants must understand and accept analytical foundations. Black-box classification systems, regardless of technical sophistication or benchmark performance, struggle to gain traction in policy contexts because stakeholders cannot verify that classifications align with their own document understanding or audit whether systematic biases affect results.

Our HTML report infrastructure, which generates publication-ready documentation available in the HTML_Reports_Fixed_Advanced_NLP directory, operationalises this transparency principle by providing stakeholder-accessible evidence that can be shared in policy briefs, included in FABLE country reports, or distributed during stakeholder consultation workshops. Each report displays not only classification results but the specific textual evidence supporting those classifications, enabling readers to form independent judgements about whether detected commitments genuinely reflect document emphasis or represent over-interpretation of marginal content. This approach operationalises the "human-in-the-loop" principle articulated by Jagannatha and Yu [8], who demonstrate that calibrated confidence scores combined with evidence transparency substantially improve expert uptake of automated classification systems by giving practitioners tools to efficiently review, challenge, and refine computational outputs rather than accepting or

rejecting them wholesale. For FABLE applications specifically, this transparency proves essential because country teams possess deep knowledge of national policy contexts that computational systems cannot capture, enabling them to identify when classifications miss important contextual factors or when medium-confidence detections warrant upgrading based on additional information not captured in analysed text.

5.3 Limitations and Future Research Directions

5.3.1 Current System Limitations

Despite substantial improvements over initial prototype implementations, our system retains important limitations that constrain its applicability and necessitate continued methodological development. Context sensitivity remains a fundamental challenge: the system detects keywords and phrases without fully understanding whether they appear in affirmative, negative, hypothetical, or counterfactual contexts. For example, text stating "biodiversity loss must be prevented" versus "biodiversity loss is projected to continue" both trigger biodiversity detection through the shared keyword "biodiversity loss," yet the first represents commitment while the second describes feared outcomes. Current keyword matching and phrase detection methods cannot reliably distinguish these cases, potentially causing the system to over-classify documents that extensively discuss problems to avoid as if they represent commitments to solutions. Future work should incorporate negation detection algorithms and causal-relationship parsing to distinguish affirmative commitments from problem descriptions.

Semantic ambiguity presents related challenges around conditional statements and contingent commitments: the system demonstrates limited understanding of policy language expressing intentions dependent on circumstances, such as "emission targets will be strengthened if economic conditions allow" or "biodiversity funding may increase subject to budget availability." These conditional commitments appear throughout policy documents as governments balance aspirational goals against fiscal and political constraints, yet current detection methods treat them equivalently to unconditional commitments. Advanced transformer models like ClimateBERT [4], which demonstrated substantial improvements in climate-relevant text classification through domain-adaptive pretraining on climate literature, could potentially improve conditional reasoning through fine-tuning on policy corpora annotated for conditionality and commitment strength.

Domain specificity constitutes another significant limitation: the system has been optimised specifically for European Union policy language and English-language documents, incorporating vocabulary, phrase patterns, and syntactic structures characteristic of EU institutional discourse. Extending to other regions—including African Union policies, ASEAN regional frameworks, or Latin American national strategies—requires substantial vocabulary adaptation and validation in new contexts where policy discourse employs different terminologies, emphasises alternative framings, and reflects distinctive institutional cultures. Similarly, application to non-English documents necessitates either development of language-specific detection systems or evaluation of multilingual transformer models that claim cross-lingual transfer capabilities.

Quantitative precision represents a conceptual rather than technical limitation: confidence scores produced by our system function as relative ordinal measures suitable for classification ranking (High confidence exceeds Medium confidence exceeds Low confidence) but do not represent calibrated probabilities suitable for direct use in probabilistic scenario analysis. While this ordinal interpretation proves appropriate for the classification task, it limits integration with Monte Carlo simulation frameworks that require probability distributions over parameters. Country teams seeking to incorporate classification uncertainty into quantitative scenarios cannot treat a 74.2% confidence score as implying 74.2% probability that a commitment exists, as the score reflects ensemble method aggregation rather than Bayesian posterior probability estimation.

Temporal dynamics constitute the final major limitation of current implementation: the system performs cross-sectional analysis treating documents as static objects rather than tracking how policy commitments evolve over time through successive versions, amendments, and implementation updates. Understanding commitment trajectories—whether they strengthen, weaken, or shift focus over time—requires temporal modelling extensions that track policy lineage and identify substantive changes across versions. This temporal dimension proves crucial for FABLE modelling, as pathway scenarios should ideally reflect current policy direction and momentum rather than treating commitments as static features of the policy landscape.

5.3.2 Integration with Advanced NLP Architectures

Future research should systematically explore integration with state-of-the-art transformer model architectures while maintaining the conservative scoring principles and evidence transparency that enable policy application. BERT fine-tuning represents the most straightforward extension: training commitment-specific BERT classifiers on manually labelled policy corpora could substantially improve classification performance by learning representations optimised for commitment detection rather than relying on general-purpose embeddings. Previous work by Zhao and colleagues [3] achieved 0.86 F1 score on policy instrument classification using BERT fine-tuned on approximately 1,700 manually annotated samples, suggesting that similar performance may be achievable for commitment detection with comparable training data. FABLE country teams could collaboratively develop this training corpus by each labelling 200 to 300 policies from their national contexts across the six commitments, creating a dataset spanning diverse policy systems and institutional contexts that would enable robust cross-regional model training.

ClimateBERT extension offers another promising direction: adapting the domain-adaptive pretraining approach pioneered for climate text to other FABLE commitments could substantially improve detection accuracy by learning commitment-specific language patterns and semantic relationships. Training a "BiodiversityBERT" through continued pretraining on scientific literature from ecology journals, Convention on Biological Diversity national reports, protected area management plans, and IPBES assessment chapters would create representations specifically optimised for biodiversity-relevant text, potentially capturing subtle distinctions between genuine conservation commitments and peripheral environmental mentions that current methods miss. Similar domain-adaptive pretraining could extend to food security using agricultural research literature and FAO reports, water management using hydrology journals and water resource plans, and other commitment domains.

Multi-label classification architectures could address a fundamental limitation of current single-commitment detection: the system treats each commitment independently despite substantial empirical evidence of systematic co-occurrence patterns. Climate-biodiversity nexus policies, food-water linkages in agricultural strategies, and economy-food interactions in rural development frameworks all represent multi-commitment integration that single-label classification misses. Multi-label models trained to simultaneously predict multiple commitments could capture these co-occurrence patterns, potentially improving detection of integrated sustainability strategies while reducing false negatives when commitments appear primarily through their relationships to other commitments rather than through direct terminology.

Zero-shot learning approaches using large language models represent the most radical architectural alternative: foundation models like GPT-4 and Claude demonstrate impressive capabilities for text classification without task-specific training data through carefully designed prompts. Comparative evaluation of zero-shot classification against our rule-based system would test whether foundation models achieve comparable precision with substantially less engineering effort, potentially enabling rapid extension to new commitment categories or policy domains without requiring extensive vocabulary curation and pattern development. However, such comparison must rigorously assess precision-recall trade-offs, evidence transparency, and operational reliability rather than focusing solely on aggregate accuracy metrics, as foundation models may

exhibit different failure modes than rule-based approaches.

5.3.3 Cross-Linguistic and Cross-Regional Extensions

Expanding beyond European Green Deal policies to support global FABLE operations requires methodological adaptations addressing linguistic diversity, institutional variation, and cultural context. Multilingual transformer models, particularly mBERT and XLM-RoBERTa supporting over 100 languages through shared multilingual representations, offer technical infrastructure for cross-lingual transfer. Initial validation should test these models on French, German, Spanish, and Portuguese policy documents within the European context, where ground truth can be established through comparison with English translations and expert validation from native-speaking country team members. Successful cross-lingual transfer within Europe would provide confidence for extending to other language families, including Swahili and Amharic for African contexts, Bahasa Indonesia and Thai for ASEAN regions, and indigenous languages for Latin American applications.

Domain adaptation proves equally important as linguistic transfer: policy language varies substantially by institutional context even when translated to common language. African Union policies characteristically emphasise poverty reduction and sustainable development framed through African development aspirations, ASEAN documents focus on resilience and regional cooperation reflecting the organisation's consensus-building approach, and Latin American national policies frequently highlight indigenous rights and agroecology concepts reflecting distinctive political-economic contexts. Direct application of European-trained vocabulary and pattern libraries to these alternative contexts would likely produce systematic biases over-weighting concepts emphasised in European discourse while missing regionally distinctive framings. Regional adaptation requires working with FABLE country teams to develop context-appropriate taxonomies through iterative refinement cycles where initial European-derived systems are validated against expert judgement and adjusted based on observed failures.

Cultural contextualisation extends beyond vocabulary to conceptual interpretation: sustainability concepts carry culturally specific meanings shaped by local ecological conditions, historical development trajectories, and prevailing value systems. "Food security" in European contexts typically emphasises nutritional quality and environmental sustainability of production systems in regions where caloric adequacy is broadly assured, while Sub-Saharan African contexts necessarily emphasise production quantity and household access to sufficient calories alongside sustainability considerations. Similarly, "biodiversity" in tropical regions rich in endemic species carries different implications than in temperate ecosystems where restoration of degraded landscapes dominates conservation discourse. Working with regional FABLE teams to develop context-appropriate interpretations ensures that classifications remain culturally valid and that confidence scores reflect region-specific policy emphasis rather than imposing European conceptual frameworks universally.

5.3.4 Integration with Other Sustainability Frameworks

The methodology developed for FABLE commitment detection generalises naturally to other sustainability frameworks requiring systematic policy mapping, offering opportunities to build integrated analytical infrastructure serving multiple assessment needs. Extension to the Sustainable Development Goals represents the most obvious application: building on previous work by Matsui and colleagues [9] and Koundouri and colleagues [10], our conservative scoring approach could improve SDG classification reliability for the 17 goals and 169 specific targets. Current SDG classification systems often suffer from over-classification due to the breadth and interconnection of SDG concepts, making nearly any policy relevant to multiple goals. Conservative scoring with evidence requirements could reduce false positives while maintaining sensitivity to genuine multi-goal integration.

Planetary boundaries framework mapping would enable assessment of policy alignment with nine Earth-system boundaries defining safe operating space for humanity, including climate change, biosphere integrity, biogeochemical flows, land-system change, freshwater use, and atmospheric aerosol loading. This application requires developing boundary-specific vocabularies and validation against scientific literature defining boundary thresholds and policy implications. Integration with FABLE commitment detection would prove particularly natural given conceptual overlaps between boundaries and commitments.

Doughnut Economics framework, which combines planetary boundaries with social foundations defining minimum standards for human wellbeing, offers another extension opportunity. Mapping policies to both ecological ceiling and social foundation dimensions would enable comprehensive assessment of whether policy portfolios balance environmental protection with human development. This application could support national and municipal governments adopting Doughnut Economics frameworks to evaluate whether existing policies adequately address both dimensions or concentrate excessively on environmental or social objectives.

IPBES Nature's Contributions to People framework provides yet another mapping target: analysing how policies address regulating contributions (like climate regulation and pollination), material contributions (like food and water provision), and non-material contributions (like cultural and spiritual values). This framework proves particularly relevant for FABLE analyses given IPBES's emphasis on integrated assessment connecting ecosystem services to human wellbeing across multiple scales.

5.3.5 Operationalisation in Policy Workflows

Moving from research demonstration to operational infrastructure integrated into FABLE country team workflows requires addressing several implementation challenges around automation, temporal tracking, stakeholder engagement, and system integration. Continuous monitoring infrastructure should establish automated pipelines that track new policy publications through government websites, parliamentary databases, and official journals, generating alerts when major documents addressing FABLE commitments appear. This monitoring proves essential for maintaining currency in rapidly evolving policy landscapes where major strategies can emerge suddenly in response to crises, as REPowerEU demonstrated with its rapid development following the Ukraine energy shock. Alert systems should provide customisable notification thresholds enabling country teams to receive immediate alerts for high-confidence detections while reviewing medium and low confidence results through periodic summary reports.

Temporal tracking capabilities should implement version control for policies, maintaining lineage information that enables tracking commitment evolution across successive iterations from initial proposals through parliamentary amendments to final regulations and subsequent implementation updates. Understanding whether commitments strengthen, weaken, or shift focus over time proves crucial for scenario development and for evaluating whether National Commitments pathways should reflect initial policy announcements or current implementation realities. Temporal analysis could also identify leading indicators of policy change by detecting shifts in discourse emphasis before formal policy revisions occur.

Stakeholder dashboards should provide interactive visualisations enabling FABLE country teams to explore policy-commitment mappings through filtering by confidence levels, commitment categories, policy types, and temporal periods. Export functionality should generate evidence briefs formatted for inclusion in country reports, stakeholder presentations, and consultation documents. Dashboard design should particularly emphasise evidence transparency by enabling users to drill down from aggregate statistics to individual classifications to specific textual evidence, supporting the iterative refinement process where country teams validate automated classifications against their policy expertise.

API integration should expose system functionality through RESTful interfaces enabling programmatic access from the FABLE Calculator and other modelling tools, facilitating automated

scenario generation where pathway assumptions dynamically incorporate latest policy classifications. API design should support both synchronous requests for real-time classification of new policy documents and batch processing for systematic corpus analysis. Authentication and rate limiting should prevent abuse while enabling legitimate research applications by academic and civil society users beyond the core FABLE network.

Feedback loop mechanisms should establish structured processes enabling country teams to validate or correct classifications, with corrections improving training data for future system iterations. This human-in-the-loop approach recognises that automated systems inevitably produce errors requiring expert correction, but that these corrections constitute valuable training signals for incremental improvement. Feedback collection should minimise country team burden by integrating with existing review workflows rather than requiring separate validation exercises, potentially through review interfaces embedded in dashboard applications or email-based correction submission.

5.4 Broader Implications for Computational Sustainability Science

Our work exemplifies a crucial maturation of computational sustainability science from proof-of-concept demonstrations establishing technical feasibility toward production-ready tools serving real policy needs with reliability and transparency standards appropriate for consequential applications. This maturation reflects broader disciplinary evolution as the field moves beyond showing that computational methods can in principle address sustainability problems toward demonstrating that they reliably do so in practice under operational conditions. Several broader lessons emerge from this transition with implications extending beyond the specific application to FABLE commitment detection.

5.4.1 Precision Over Recall in Policy Contexts

The machine learning research community conventionally optimises F1-score as the primary performance metric, treating precision and recall symmetrically through their harmonic mean. This symmetry reflects an implicit assumption that false positives and false negatives carry equivalent costs, an assumption that holds for many technical applications but fails systematically in policy contexts where error consequences diverge sharply. False positives suggesting commitments where none exist can trigger misallocation of conservation funding toward interventions lacking governmental support, misalignment of modelling assumptions with policy realities leading to scenarios disconnected from plausible implementation pathways, and stakeholder confusion about governmental intentions that undermines trust in analytical processes. Our achievement of greater than 90% false-positive reduction versus baseline approaches demonstrates that conservative precision—deliberately accepting lower recall to prevent misleading classifications—better serves decision-making in contexts where false positives carry genuine consequences for resource allocation and strategy development.

This precision-prioritising design philosophy requires resisting the momentum of mainstream machine learning research emphasising recall improvements and aggregate performance gains measured through benchmark datasets. In policy applications, the most valuable system is not necessarily the one detecting the most commitments but rather the one whose detections can be trusted sufficiently that stakeholders act on them confidently. A system achieving 95% precision at 70% recall proves more valuable than one achieving 85% precision at 90% recall if the precision difference determines whether country teams trust results sufficiently to incorporate them into quantitative scenarios. Future research in computational sustainability should explicitly optimise for application-appropriate metrics reflecting actual usage contexts rather than uncritically adopting performance measures from general machine learning literature.

5.4.2 Transparency Builds Trust

Black-box deep learning models, despite often achieving superior technical performance on held-out test sets compared to interpretable alternatives, struggle to gain traction in policy applications because stakeholders cannot audit reasoning processes or verify that classifications align with their domain understanding. This trust barrier proves particularly severe in contentious policy domains where different actors hold competing visions of appropriate sustainability pathways and where analytical outputs influence resource allocation across those competing visions. Our mandatory evidence trails providing context excerpts, keyword occurrence counts, and method contribution breakdowns enable non-technical experts to validate classifications and form independent judgements about their validity, fostering trust essential for system adoption in collaborative modelling processes like FABLE Scenathons where diverse stakeholders must collectively accept analytical foundations.

The transparency imperative extends beyond technical implementation to research communication: papers describing policy NLP systems should foreground evidence and examples rather than primarily reporting aggregate performance statistics. Demonstrating that the system correctly classifies the EU Energy Integration Strategy with 74.2% climate confidence based on 67 keyword matches provides more persuasive validation than reporting 0.89 F1-score on a held-out test set, because the former enables readers to verify correctness through their own document understanding while the latter requires trust in dataset construction and evaluation protocols. Future work on explainable AI for sustainability should prioritise evidence transparency and stakeholder interpretability over marginal accuracy gains, recognising that the most technically sophisticated system proves useless if stakeholders reject its outputs as incomprehensible black-box classifications.

5.4.3 Domain Expertise Remains Central

While natural language processing automates labour-intensive analysis that would otherwise require hundreds of person-hours of manual document review, domain expertise from sustainability science proves indispensable throughout the analytical pipeline from system design through validation to application. Vocabulary curation requires understanding which terms sustainability scientists and policymakers actually use to discuss commitments, pattern development depends on recognising characteristic policy discourse structures, confidence threshold setting demands judgement about acceptable precision-recall trade-offs in application contexts, and classification interpretation necessitates understanding policy contexts that computational systems cannot extract from text alone. Successful computational sustainability science requires genuine interdisciplinary collaboration where domain experts and computational specialists jointly design systems, validate outputs, and interpret results, rather than merely applying computer science methods to sustainability domains as technical service provision.

This centrality of domain expertise contradicts common narratives suggesting that artificial intelligence will replace expert judgement with automated analysis. Our experience developing FABLE commitment detection reveals instead that automation shifts rather than eliminates expertise requirements: instead of reading every document manually to identify commitments, experts curate vocabularies, validate automated classifications, interpret medium-confidence cases requiring contextual judgement, and translate classification results into modelling assumptions. These tasks demand deeper engagement with policy content than simple document screening, requiring experts who understand both commitment concepts and institutional contexts rather than research assistants following mechanical coding protocols. Future systems should therefore emphasise expert augmentation rather than replacement, designing human-in-the-loop workflows that leverage automation for scalable screening while preserving space for expert judgement on substantive interpretive questions.

5.4.4 From Research to Practice

Academic NLP research conventionally emphasises novel architectures evaluated on standardised benchmark datasets under controlled conditions optimising for aggregate performance metrics. Operational policy tools serving real stakeholder needs require fundamentally different priorities: reliability under diverse conditions, computational scalability for large document corpora, system maintainability as policy language evolves, result transparency enabling stakeholder validation, and practical integration with existing analytical workflows. Our emphasis on production-ready implementation achieving 100% processing success rate across all 42 European Green Deal documents, completing batch processing in under 5 minutes, generating automated HTML reports accessible to non-technical users, and maintaining evidence trails for every classification exemplifies engineering rigour complementing research innovation. This engineering focus represents neither methodological conservatism nor technical limitation but rather recognition that systems achieving 98% success rates prove unsuitable for operational deployment where processing failures require manual intervention, while systems demanding computational resources beyond typical country team infrastructure remain inaccessible despite impressive benchmark performance.

The path from research prototype to operational tool involves addressing numerous unglamorous implementation challenges including robust error handling, efficient processing pipelines, user-friendly interfaces, comprehensive documentation, and automated testing frameworks. These engineering investments receive limited recognition in academic publication venues emphasising methodological novelty but prove essential for real-world deployment and sustained use. Future computational sustainability science should develop evaluation frameworks recognising operational readiness as a distinct contribution worthy of scholarly recognition, potentially through demonstration tracks at conferences, repositories highlighting production-ready implementations, and publication venues valuing engineering contributions alongside algorithmic innovations.

6 Conclusions and Recommendations

This paper presented a production-ready multi-method NLP system for informing the building blocks that shape FABLE Pathways utilizing European Green Deal policy documents through rigorous evidence-based classification. The system integrates four complementary analytical techniques—enhanced keyword analysis with weighted vocabulary categories, spaCy phrase detection capturing multi-word expressions, TF-IDF semantic similarity identifying documents genuinely focused on commitment domains, and syntactic pattern matching recognising policy-specific discourse structures—combined through conservative ensemble scoring that prioritises precision over recall. This architectural design achieved greater than 90% reduction in false positive classifications compared to baseline approaches while maintaining high sensitivity to genuine policy content, demonstrating that methodological rigour substantially improves system reliability for policy-facing applications where misleading classifications carry consequences for decision quality and resource allocation.

Corpus-level analysis of 42 European Green Deal strategy documents revealed systematic patterns in how major EU policies address the tenets that shape the FABLE Pathways. Climate mitigation dominates the policy landscape with high confidence detection in 43% of documents, reflecting the European Green Deal's foundational emphasis on climate neutrality by 2050 and emission reduction targets forming the organising framework for sectoral transformation. Biodiversity integration appears more selectively, achieving medium-to-high confidence in 55% of policies through successful mainstreaming of the EU Biodiversity Strategy 2030, though remaining more sector-specific than the nearly universal climate consideration. Food security and economy commitments concentrate primarily in sectoral agricultural and rural development policies rather than pervading cross-cutting frameworks, suggesting potential disconnection between

sectoral strategies and overarching sustainability visions. Water management and fertiliser use exhibit the weakest corpus presence with high-confidence detection in only 5% and 2% of documents respectively, indicating severe policy gaps requiring explicit attention in sustainability planning and potentially creating misalignment between stated environmental objectives and implementation mechanisms.

Case studies of the REPowerEU Energy Plan and EU Energy Integration Strategy demonstrated the system's ability to capture nuanced policy positions through graduated confidence classifications. REPowerEU's medium confidence classification at 35.3% appropriately captures the document's complex relationship to climate commitments, where renewable energy acceleration and fossil fuel phase-out deliver genuine climate benefits but serve primarily energy security objectives rather than emission reduction per se. The EU Energy Integration Strategy's high confidence climate classification at 74.2% based on 67 strong keyword matches correctly identifies decarbonisation as the foundational driver of energy system transformation, while medium and low confidence classifications for economy and biodiversity accurately distinguish secondary considerations from primary policy emphasis. These case studies validate the proposition that conservative evidence-based scoring produces classifications suitable for informing quantitative scenario development and policy analysis, avoiding both the over-classification failures of aggressive detection systems and the excessive caution of approaches missing genuine policy commitments.

6.1 Strategic Recommendations for Sustainability Scientists

Sustainability scientists seeking to develop or deploy natural language processing systems for policy analysis should adopt several strategic principles emerging from our research and operational experience. Conservative NLP architectures prioritising precision over recall prove essential for policy-facing applications where false positive classifications can mislead decision-makers and misallocate resources, even when this conservatism accepts lower recall and misses some genuine commitments. The precision-recall trade-off should explicitly reflect application context and error consequences rather than uncritically optimising F1-scores following conventions from general machine learning literature. Researchers should begin with existing validated frameworks rather than developing classification systems from scratch, leveraging resources like the OSDG toolkit for Sustainable Development Goal classification, ClimateBERT for climate-relevant text analysis, and our FABLE-NC system for commitment detection, then systematically validating these frameworks on their specific document corpora to assess performance and identify adaptation requirements.

High-quality training data proves essential for both validating rule-based systems and supporting fine-tuning of transformer models: sustainability scientists should invest resources in developing manually annotated policy corpora comprising 200 to 500 documents per commitment category, ensuring annotation quality through multiple independent coders and inter-rater reliability assessment. This training data investment pays dividends through improved system performance, rigorous validation capability, and foundation for future machine learning extensions. Early engagement with computational experts ensures that methodological choices align with domain research questions rather than imposing computational convenience, fostering genuine interdisciplinary collaboration where sustainability knowledge and computational capabilities jointly shape system design.

Evidence transparency should be treated as a non-negotiable requirement rather than optional enhancement: all classification systems should provide context excerpts showing actual textual evidence, keyword occurrence lists enabling audit of detection triggers, and method contribution breakdowns revealing how different analytical components influenced final classifications. This transparency infrastructure enables domain experts lacking computational training to validate results, identify systematic errors, and develop trust essential for system adoption. Publication of classification systems should foreground evidence and concrete examples rather

than primarily reporting aggregate performance statistics, recognising that stakeholders evaluate trustworthiness through their ability to verify specific classifications against their domain understanding rather than through confidence in held-out test set performance metrics.

6.2 Operational Guidance for FABLE Country Teams

FABLE country teams integrating automated commitment detection into their pathway development workflows should adopt confidence-stratified interpretation protocols matching detection certainty to modelling applications. High-confidence detections achieving 60% or greater scores provide sufficiently reliable signals for direct incorporation into quantitative National Commitments pathway scenarios, justifying specific parameter values and policy implementation assumptions reflecting governmental dedication to detected commitments. Medium-confidence detections spanning 35% to 60% should trigger stakeholder consultation and expert review rather than direct scenario incorporation, serving as flags identifying policies requiring interpretive judgement about whether genuine commitments exist despite mixed textual signals or whether apparent commitment language reflects aspirational framing rather than actionable policy dedication.

Vocabulary and pattern adaptation to national language and institutional context substantially improves system performance beyond direct application of European-derived detection systems. Country teams working in non-English contexts or outside European institutional frameworks should systematically validate detection systems against expert judgement, documenting cases where classifications fail to align with policy understanding, and iteratively refining vocabularies and patterns to capture regionally distinctive policy discourse. This adaptation process should engage team members with deep knowledge of national policy traditions and institutional language conventions, recognising that effective commitment detection requires understanding of both sustainability concepts and policy communication practices.

Collaborative development of labelled policy repositories across the global FABLE network would substantially improve system performance for all participating country teams while distributing annotation effort across multiple groups. Each country team contributing labelled policies from their national contexts creates training datasets spanning diverse policy systems, institutional traditions, and linguistic patterns, enabling development of robust classification models that generalise across contexts rather than overfitting to particular regional characteristics. This collaborative approach transforms annotation from individual country burden into collective infrastructure investment with shared benefits.

Expert qualitative analysis should complement rather than be replaced by automated screening, particularly for medium-confidence cases where textual signals prove ambiguous. Country teams should view NLP systems as tools for efficient preliminary screening that directs expert attention toward policies likely addressing commitments, rather than as authoritative classification mechanisms eliminating need for expert interpretation. This human-in-the-loop approach combines computational scalability with expert judgement, leveraging automation for breadth while preserving human insight for depth.

6.3 Research Priorities for NLP Scientists

Natural language processing researchers working on sustainability applications should prioritise development of policy-specific benchmark datasets with rigorous annotation protocols, inter-rater reliability assessment, and documentation of annotation decisions enabling others to understand and replicate classification criteria. Current policy NLP research suffers from fragmented evaluation across incompatible datasets with unclear annotation quality, preventing meaningful comparison of methods and hindering cumulative progress. Standardised benchmarks covering major policy domains—climate mitigation, biodiversity conservation, food security, water

management, sustainable development goals—would accelerate methodological innovation by enabling systematic evaluation of alternative approaches under controlled conditions.

Conservative ensemble architectures employing weighted voting, calibrated confidence scoring, and explicit precision-recall optimisation deserve sustained research attention as alternatives to single-model approaches dominating current literature. While transformer models achieve impressive benchmark performance, policy applications require reliability guarantees and error transparency that ensemble methods can provide through method triangulation and evidence validation. Research should systematically compare ensemble approaches against individual models on policy classification tasks, evaluating not only aggregate accuracy but also precision-recall trade-offs, evidence transparency, failure mode characteristics, and operational reliability under diverse conditions.

Domain-adaptive pretraining beyond climate toward other sustainability commitments represents a promising research direction extending the successful ClimateBERT approach. Training commitment-specific language models through continued pretraining on domain literature—biodiversity models trained on ecology journals and conservation reports, food security models trained on agricultural research and nutrition literature, water management models trained on hydrology journals and resource management plans—would create representations optimised for detecting commitment-relevant policy discourse. This approach could substantially improve classification performance while maintaining interpretability through vocabulary-based explanations connecting model predictions to recognised domain terminologies.

Temporal policy analysis tracking commitment language evolution across document versions and forecasting policy trajectories remains largely unexplored despite clear application relevance. Research developing methods to compare successive policy versions, identify substantive commitment changes versus editorial refinements, and project future policy directions based on discourse trends would enable more sophisticated pathway scenario development incorporating policy momentum and direction rather than treating commitments as static. This temporal dimension proves particularly important for rapidly evolving domains like climate policy where major strategies undergo frequent revision in response to scientific updates and political developments.

Explainable policy NLP tools foregrounding evidence over prediction confidence deserve sustained development effort recognising that adoption in policy contexts depends on transparency and interpretability rather than marginal accuracy gains. Research should explore interfaces presenting classifications through ranked evidence lists showing strongest supporting textual passages, method contribution breakdowns revealing analytical reasoning, and counterfactual explanations demonstrating how classification would change if particular textual features were altered. This human-centred design approach treats classification systems as decision support tools augmenting expert judgement rather than black-box automation replacing human interpretation.

6.4 Vision for Integrated Computational Sustainability Infrastructure

The integration of advanced natural language processing with participatory modelling frameworks like FABLE represents a crucial frontier in computational sustainability science, enabling evidence-based policy analysis at scales and speeds previously impossible through manual methods alone. As transformer architectures continue improving, multilingual capabilities expand, and domain-adaptive pretraining techniques mature, automated policy analysis will evolve from specialised research application into standard infrastructure supporting sustainability planning worldwide. This infrastructural vision encompasses not merely technical capability but institutional embedding where computational tools become routine components of policy development processes, scenario analysis workflows, and stakeholder engagement activities.

Yet technology alone cannot achieve sustainable transformation, regardless of computational sophistication or analytical power. Natural language processing tools serve sustainability ob-

jectives most effectively when embedded in genuine stakeholder engagement processes, iterative refinement cycles incorporating expert feedback, and institutional learning systems that improve analytical capabilities through operational experience. The FABLE Scenathon methodology—bringing together researchers, policymakers, civil society representatives, and private sector stakeholders in collaborative modelling exercises developing alternative pathway scenarios—provides ideal context for deploying and refining computational tools that support rather than replace human deliberation. Automated commitment detection accelerates preliminary policy screening, enables rapid update of scenarios as policies evolve, and generates evidence briefs documenting governmental positions, but substantive interpretation, stakeholder validation, and translation into modelling assumptions necessarily remain human activities requiring judgement that algorithms cannot provide.

Our aspirational vision encompasses a future where every FABLE country team employs validated natural language processing systems systematically tracking national policy commitments, automatically updating pathway scenario assumptions as new policies emerge, and generating evidence briefs supporting stakeholder consultations and policy dialogues. This computational infrastructure would enable rapid scenario development responding to policy announcements within days rather than months, comparative cross-country analysis identifying divergent commitment patterns and enabling peer learning, and early warning systems detecting potential misalignments between stated commitments and implementation mechanisms before they crystallise into policy failures. The infrastructure would support not merely technical analysis but democratic deliberation by making policy content accessible through structured summaries, evidence trails, and temporal tracking revealing how commitments evolve through political processes.

Achieving this vision by 2030 requires sustained interdisciplinary collaboration bridging computer science, sustainability science, policy analysis, and participatory modelling communities. Technical advances in natural language processing provide necessary foundation but prove insufficient without complementary advances in institutional adoption, stakeholder engagement practices, and integration with decision-making workflows. The path forward demands continued methodological innovation developing more reliable and transparent classification systems, expanded validation across diverse policy contexts and linguistic environments, collaborative infrastructure development creating shared resources serving the global sustainability science community, and patient institutional engagement building trust and demonstrating value through concrete applications addressing real stakeholder needs. Our work on FABLE commitment detection represents one step along this path, demonstrating technical feasibility and operational viability while identifying challenges requiring sustained attention from researchers, practitioners, and funders committed to building computational infrastructure supporting evidence-based sustainability transformation.

Acknowledgments

The authors thank the FABLE Consortium participants in the FABLE Consortium Meeting in Istanbul in June 2025 for their feedback, the European Green Deal policy team for document compilation, and the broader FABLE Greece team for validation feedback. This work builds on open-source NLP libraries (spaCy, scikit-learn) and benefits from the global sustainability science community's commitment to transparent, reproducible research methods.

Conflicts of Interest

The authors declare no conflicts of interest. This research received no external funding.

References

[1] FABLE Consortium (2024). *Pathways to Sustainable Land-Use and Food Systems*. In J.D. Sachs, G. Lafortune, & G. Fuller (Eds.), *The SDGs and the UN Summit of the Future: Sustainable Development Report 2024* (Part 4). Paris: Sustainable Development Solutions Network (SDSN); Dublin: Dublin University Press. DOI: 10.25546/108572.

[2] Jin, Z., & Mihalcea, R. (2023). Natural Language Processing for Policymaking. In *Handbook of Computational Social Science for Policy* (pp. 141–162). Springer, Cham. DOI: 10.1007/978-3-031-16624-2_7.

[3] Zhao, J., & Li, C. (2022). Research on the Classification of Policy Instruments Based on BERT Model. *Discrete Dynamics in Nature and Society*, 2022, Article 6123348. DOI: 10.1155/2022/6123348.

[4] Webersinke, N., Kraus, M., Bingler, J.A., & Leippold, M. (2022). ClimateBERT: A Pretrained Language Model for Climate-Related Text. In *Proceedings of the AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*. Arlington, VA. arXiv:2110.12010.

[5] Schimanski, T., Bingler, J.A., Kraus, M., Hyslop, C., & Leippold, M. (2023). ClimateBERT-NetZero: Detecting and Assessing Net Zero and Reduction Targets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)* (pp. 15745–15756). Singapore: Association for Computational Linguistics. DOI: 10.18653/v1/2023.emnlp-main.975.

[6] Smith, T.B., Vacca, R., Mantegazza, L., & Capua, I. (2021). Natural language processing and network analysis provide novel insights on policy and scientific discourse around Sustainable Development Goals. *Scientific Reports*, 11, Article 22427. DOI: 10.1038/s41598-021-01801-6.

[7] Jang, H.-S. (2025). Ensemble Learning Model for Industrial Policy Classification Using Automated Hyperparameter Optimization. *Electronics*, 14(20), Article 3974. DOI: 10.3390/electronics14203974.

[8] Jagannatha, A., & Yu, H. (2020). Calibrating Structured Output Predictors for Natural Language Processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)* (pp. 2078–2092). Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.188.

[9] Matsui, T., Suzuki, K., & Ando, K. (2022). A natural language processing model for supporting sustainable development goals: Translating semantics, visualizing nexus, and connecting stakeholders. *Sustainability Science*, 17, 969–985. DOI: 10.1007/s11625-022-01093-3.

[10] Koundouri, P., Aslanidis, P.-S., Dellis, K., Plataniotis, A., & Feretzakis, G. (2025). Mapping human security strategies to sustainable development goals: A machine learning approach. *Discover Sustainability*, 6, Article 96. DOI: 10.1007/s43621-025-00883-w.